UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE GEOGRAFÍA E HISTORIA

TESIS DOCTORAL

Oportunidades de los datos geolocalizados de Twitter en el estudio de la movilidad metropolitana

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Joaquín Osorio Arjona

Director

Juan Carlos García Palomares

Madrid, 2020

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE GEOGRAFÍA E HISTORIA

TESIS DOCTORAL

Oportunidades de los datos geolocalizados de Twitter en el estudio de la movilidad metropolitana

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Joaquín Osorio Arjona

DIRECTOR

Juan Carlos García Palomares

"A journey of a thousand miles begins with a single step."

Lao-Tze

ABSTRACT

The growth of metropolitan areas and the specialization of urban zones into residential or labor districts have led to an increase in the number, length and duration of daily trips. This raises the need for a constant stream of data to provide updated information on the spatial and temporal characteristics of metropolitan mobility. In recent decades, the appearance of a series of data sources based on Information and Communication Technologies have allowed to collect large amounts of data quickly, frequently, and in some cases, at low costs. These data tend to have high spatial and temporal resolution, are easy to update, and can provide almost real-time monitoring of metropolitan mobility patterns. Some of the most important new data sources are social networks, platforms where internet users communicate and share ideas, opinions, and information.

The objective of this doctoral thesis is to determine the extent to which data created using new Big Data-based technologies can be used as an alternative to traditional data sources to obtain information related to metropolitan mobility. Specifically, this thesis has used data from the social network *Twitter*, which provides high volumes of free, updated, high space-time resolution data. The possibility of geotagging tweets and their geometry as point features means that these data can be easily processed, analyzed and visualized in Geographic Information Systems.

This thesis proposes to analyze the mobility patterns of the Madrid Metropolitan Area from various case studies, taking advantage of the spatial, temporal, and semantic value of the *Twitter* data. For that purpose, a method for storing and processing *Twitter* data has been devised, by creating a series of filters for selecting valid *Twitter* users. The data were then enriched with additional information, such as land registry data based on the location of the *tweets*. Finally, the thesis has created a specific methodology for analyzing and visualizing data from each case study.

In the first case study, the thesis has visualized home-work travel flows by creating an origin-destination matrix with which to analyze movement between the different metropolitan zones. In the second case study, the thesis has designed space-time trajectories to represent the individual mobility of *Twitter* users in certain areas of the city. In the third case study, the thesis has analyzed the influence of factors such as location or mode of transportation on mobility that university campuses attract. In the fourth case study, the thesis has used *Twitter* data to visualize the space-time distribution of the population during a macro event, and to obtain the origin place of each visitor.

Finally, in the fifth case study the thesis has analyzed the user opinions of a public transport system such as the Madrid Metro based on the texts from the *tweets* and the variables that affect the spatial distribution of these opinions.

The results obtained have shown that the geographical distribution of *Twitter* users and home-work travel flows are generally consistent with the data provided by official sources, although the number of users found in the center of Madrid is overestimated. It has been observed at different scales a predominance of centripetal travel flows originating in the outskirts or suburbs and moving to the central areas of the city. By creating 2D and 3D space-time trajectories it was able to show that a city residential area tends to generate travel flows, while other areas in which business or study activity predominate usually attract travelers.

An analysis of the university population mobility has shown that *Twitter* users tend to live in the city center or in the outskirts or suburbs near to the campus they attend, and the influence of factors like the income level of the places of residence and the access to transport networks. An analysis of the *Twitter* fingerprint generated during a macro event showed a higher volume of users compared to a typical week, and a significant concentration of activity in the areas where the event was held. Finally, the results have shown that Metro users tend to report problems involving the busiest lines or stops that are either located in the city center or are linked with other Metro lines.

This doctoral thesis paves the way for new lines of research in the future. Despite its limitations, such as the low volume of geotagged *tweets* or the presence of demographic and socioeconomic biases, streaming and enriching data with information from different sources has increased the value and versatility of the sample data. The interoperability and flexibility of *Twitter* data will facilitate the investigation of different aspects of metropolitan mobility not addressed in this thesis.

RESUMEN

El crecimiento de las áreas metropolitanas y la especialización de zonas urbanas en áreas de trabajo o residencia han conllevado el aumento de la cantidad, longitud y duración de los viajes que se realizan diariamente. Es necesario contar con datos constantes que proporcionen información actualizada sobre las características espaciales y temporales de la movilidad metropolitana. En las últimas décadas han aparecido una serie de fuentes de datos vinculadas a las Tecnologías de la Información y la Comunicación, que permiten recoger una gran cantidad de datos de forma rápida y frecuente, y en algunos casos, a bajo coste. Estos datos suelen contar con una alta resolución espacio-temporal, son fáciles de actualizar, y permiten la monitorización de patrones de movilidad metropolitana a tiempo casi real. Entre las nuevas fuentes de datos destacan las redes sociales, plataformas donde los usuarios de internet se comunican y comparten ideas, opiniones, e información.

El objetivo de la presente tesis doctoral es evaluar la capacidad de los datos creados por las nuevas tecnologías basadas en el *Big Data* como métodos alternativos a las fuentes tradicionales de datos para obtener información relacionada con la movilidad metropolitana. En concreto, esta tesis trabaja con datos de la red social *Twitter*, debido a la posibilidad de obtener de forma gratuita un volumen alto y constante de datos con un alto detalle espacio-temporal. La posibilidad de geolocalizar los *tweets* y su geometría en forma de entidades de punto permite el fácil tratamiento, análisis y visualización de estos datos en un Sistema de Información Geográfica.

Esta tesis propone analizar los patrones de movilidad del Área Metropolitana de Madrid a partir de diversos casos de estudio, aprovechando el valor espacial, temporal, y semántico de los datos de *Twitter*. Para llevar a cabo esta investigación, se establece una metodología de almacenamiento y procesado de datos en la que se establece una serie de filtros que permitan seleccionar los usuarios válidos para el estudio. A continuación, se enriquecen los datos con información adicional como los datos de uso del suelo desde donde se publican los *tweets*. Por último, la tesis elabora una metodología específica de análisis y visualización de los datos para cada caso de estudio.

En el primer caso de estudio, la tesis propone visualizar los flujos de viajes residencia-trabajo, a través de la elaboración de una matriz Origen-Destino con la que analizar los movimientos entre las diferentes zonas del área metropolitana. El segundo caso de estudio busca representar la movilidad individual de los usuarios de *Twitter* en determinadas zonas de la ciudad, mediante el diseño de caminos espacio-temporales. En el tercer caso

de estudio, la tesis estudia la influencia de factores como la localización o el modo de transporte sobre la movilidad que atraen los campus universitarios. El cuarto caso de estudio emplea los datos de *Twitter* para visualizar la distribución espacio-temporal de la población durante un macroevento, y obtener el lugar de procedencia de los visitantes. Finalmente, el quinto caso de estudio analiza la percepción de un sistema de transporte público como el Metro de Madrid a partir de los textos de los *tweets* y las variables que afectan a la distribución espacial de estas percepciones.

Los resultados obtenidos han enseñado que la distribución geográfica de los usuarios de *Twitter* y los flujos de movilidad residencia-trabajo presentan una situación cercana a los datos suministrados por fuentes de datos oficiales, aunque con una sobreestimación de usuarios en las áreas centrales de la zona de estudio. Se ha observado a diferentes escalas una predominancia de los flujos de viaje centrípetos con origen en los municipios o distritos de la periferia, y destino en las áreas centrales del municipio de Madrid. A partir del diseño de caminos espacio-temporales tanto en 2D como en 3D se ha visualizado como una zona residencial de la ciudad tiende a generar viajes, y otras zonas con actividad predominante de trabajo o estudios suelen ser espacios de atracción de viajes.

Al analizar la movilidad universitaria, la tesis ha mostrado como los usuarios de *Twitter* tienden a residir en áreas centrales o en distritos o municipios próximos al campus al que asisten, y la influencia del nivel de renta del lugar de residencia, y del acceso a las redes del transporte. En el estudio de la huella digital generada en *Twitter* durante un evento de masas, se ha observado un mayor volumen de usuarios respecto a una semana habitual, y una concentración importante de actividad en las áreas donde se ha celebrado el evento. Por último, la tesis ha señalado que los usuarios que usan el servicio de Metro tienden a reportar problemas en las líneas más transitadas o en las paradas que están ubicadas en el centro de la ciudad o que sirven de enlace con otras líneas de Metro.

Esta tesis doctoral abre nuevas posibles líneas de investigación en el futuro. Aunque se han encontrado algunas limitaciones como un bajo volumen de *tweets* geolocalizados o la presencia de sesgos demográficos y socioeconómicos, la descarga continua de datos y el enriquecimiento a partir de información proporcionada por distintas fuentes de datos, permiten aumentar el valor y versatilidad de los datos de la muestra. La interoperabilidad y flexibilidad de los datos de *Twitter* posibilitan la investigación en distintas escalas espaciales y temporales de diferentes casuísticas relacionadas con la movilidad metropolitana no tratadas en esta tesis.

AGRADECIMIENTOS

Esta tesis doctoral culmina un viaje de diez años que inicié en 2009 cuando abandoné Antequera y me desplacé a Sevilla para iniciar el Grado de Geografía e Historia en la Universidad Pablo de Olavide. Desde siempre me habían atraído los mapas y ya tenía el objetivo de convertirme en geógrafo o cartógrafo, cuando en tercero de grado hubo una asignatura de Tecnologías de Información Geográfica que me abrió la puerta al mundo de los SIG y la cartografía. Las áreas metropolitanas son una entidad que ejercían una fuerte atracción en mí, por lo que mi trabajo de fin de grado trató sobre la coremática como método cartográfico original para analizar los flujos que se dan en el Área Metropolitana de Sevilla. En 2014 me mudé a Alcalá de Henares para estudiar el Máster de TIG. Aquí tuve la oportunidad de trabajar en otro tema que siempre me ha interesado: las redes de transporte. Mi trabajo de fin de master fue la cartografía de una nueva red de autobuses en Alcalá de Henares que fuese más eficaz que la que había por entonces.

Con mi gusto por los SIG, la cartografía, las áreas metropolitanas, y las redes de transporte en mi cabeza, me matriculé en 2015 al programa de Doctorado en Geografía por la Universidad Complutense de Madrid. Por aquel entonces, apenas estaba puesto en aspectos como el *Big Data*, o la posibilidad de usar *Twitter* como fuente de datos para elaborar mapas con los que estudiar la movilidad en áreas metropolitanas (el único uso que le tenía a *Twitter* en la mente era usar mi cuenta @cartografo87 para promocionar los mapas que hiciese en *ArcGIS*). También pensaba que acabaría realizando todo el doctorado desde mi casa en Antequera, ya que no teníamos los recursos económicos que me permitiesen vivir en Madrid por casi un lustro, y esta vez no contaba con un programa de beca-alojamiento que si tuve cuando c cursé el máster. Poco sospechaba que echar una solicitud para el programa FPU iba a cambiar tanto mi vida en estos últimos años.

Ahora que lo pienso, estos últimos años han pasado volando. Parece que fue ayer el día en el que entré en el despacho de mi tutor por primera vez para hablar del tema de mi tesis doctoral. En estos últimos cinco años he tenido muchas experiencias: publicar artículos, viajar a diversas partes de España a exponer mis avances en conferencias, dar docencia a alumnos de grado, etc. Creo que toda la experiencia que he tenido en estos años de elaboración de la tesis ha sido bastante positiva y me ha ayudado a crecer tanto académicamente como persona. Siempre he sido una persona cortada y algo tímida, por lo que tener que dar clases o realizar presentaciones me ha ayudado a ganar confianza y romper un poco el cascarón de seguridad. He aprendido bastante sobre el *Big Data* en

general y todo el enorme potencial que tiene *Twitter* en particular. Y aunque esta tesis sea el punto final a una aventura, tengo muchas ganas de ponerme ya a realizar nuevas investigaciones geográficas basadas en datos de *Twitter*.

Esta tesis no hubiera posible sin la ayuda de mi tutor Juan Carlos García Palomares, quien ha sido mi guía y mi gran compañero de aventuras en estos cinco años. Quiero agradecerle de corazón todo el tiempo y el esfuerzo que me ha dedicado, y la paciencia que ha tenido conmigo, sobre todo en esos momentos en los que me costaba sacar adelante ciertos apartados de la tesis, y mi mente se bloqueaba automáticamente haciendo que repitiese los mismos errores varias veces. Gracias a Juan Carlos he aprendido multitud de cosas que me van a servir para el futuro, y he podido dar mis primeros pasos tanto en la publicación académica como en la docencia.

A lo largo de esta tesis he compartido facultad con una serie de compañeros a los que quiero agradecer por su compañía y amistad. Primero quiero agradecer a los compañeros de doctorado o de trabajo que compartieron sitio: Borja Moya, Gustavo Romanillos, Amparo Moyano, Claudia Yubero, Chema Fernández, Néstor Campos, Carolina Barros, Julio Gómez, Elena Ferreiro, Onel Pérez, Daniela Arias, Sofía Mendoza y Kike Santiago. Y también a los compañeros que estuvieron brevemente: Henar Salas, Jaime Díaz, Gennaro Angiello, Harold Cardona, José Barros, Ofelia Martínez, Marina Barber, Oswaldo Pinillos, Toni Domenech, y Marcelo Fernandes.

Agradecer también a los compañeros del grupo de investigación tGIS (tanto los que siguen como los que ya no están) del que me siento orgulloso de haber formado parte: Javier Gutiérrez, Henar Salas, Ana Condeço, Juanjo Michellini, Enrique Pozo, Juana Moya, Marcin Stepniak, Inmaculada Mohino, Rubén Talavera, y Rocío Pérez. Quiero extender mi agradecimiento a todos los miembros del Departamento de Geografía de la Universidad Complutense (especialmente a Carmen Mínguez con quien he compartido docencia de SIG para Arqueología durante un par de cursos), y a las compañeras del grupo de informática asociado al proyecto SocialBigdata-CM, especialmente a Guadalupe Miñana y Yolanda García.

Para conseguir en el título de Doctorado la mención internacional, he realizado una estancia tres meses en la VSB Technical University de Ostrava, República Checa. Quiero agradecer especialmente a Jiri Horak por haber trabajado conmigo en el artículo con el que he cerrado esta tesis doctoral, a Igor Ivan por haberme aceptado en el departamento de geoinformática durante esos tres meses, a Michal Kacmarik por toda la ayuda que me

ha dado, y a Radek Svoboda por toda la ayuda ofertada en el análisis de textos de los *tweets*, un apartado informático que hubiera sido incapaz de haber hecho por mí mismo.

Estos cinco años han sido emocionantes, pero en algunos momentos, han conllevado en algunos momentos cierto desgaste psicológico y de energía. Si hay alguien que merece un agradecimiento muy especial, esa es mi madre María, quien siempre ha estado detrás de mí, preocupándose todo el rato de mi bienestar y salud. También agradecer al resto de mi familia que siempre se ha interesado por lo que he estado haciendo estos años, a mi abuela Antonia, a mis tíos Carlos, Toni y Luis, y a mis primas Nuria, Carmen, Marta, Sara y Raquel. Y a mi abuelo Luis, que en paz descanse.

Como ya he comentado, obtener una beca FPU ha sido vital para la elaboración de la tesis doctoral, al darme los recursos necesarios para poder vivir en Madrid, para poder optar a dar mis primeros pasos en la docencia, y para poder realizar una estancia de tres meses fuera de España. Quiero agradecer la financiación del Ministerio de Educación, Cultura y Deporte por ello (Programa FPU AP2015-0147). Finalmente, quiero agradecer también los distintos proyectos y redes en los que se han integrado mis artículos. Por tanto, agradecer al Ministerio de Economía, Industria y Competitividad (MINECO) y al Fondo de Desarrollo Regional Europeo (ERDF) (Proyecto TRA2015-65283-R), al Gobierno Regional de Madrid (SOCIALBIGDATA-CM, S2015/HUM-3427), y al Ministerio de Ciencia, Innovación y Universidades y el Fondo Regional Europeo de Desarrollo (Proyecto DynMobility, RTI2018-098402-B-I00).

Muchísimas gracias a todas las personas que me han apoyado, ayudado, o contribuyeron al desarrollo de esta tesis doctoral.

ÍNDICE DE CONTENIDO

ABSTRACT **5**

RESUMEN **7**

AGRADECIMIENTOS **9**

ÍNDICE DE CONTENIDO **13**

ÍNDICE DE TABLAS **16**

ÍNDICE DE FIGURAS **17**

1. INTRODUCCIÓN **19**

1.1. Interés y oportunidad de la investigación 20

1.2. Preguntas de investigación 27

1.3. Objetivos 30

1.4. Estructura del trabajo 37

2. MARCO TEÓRICO **41**

2.1. Fuentes tradicionales de datos 42

2.2. Nuevas fuentes de datos para el estudio de la movilidad urbana 44

2.2.1. Introducción a las nuevas fuentes de datos *44*

2.2.2. Características de las nuevas fuentes de datos *48*

2.2.3. Clasificación de las nuevas fuentes de datos *52*

2.2.4. Ventajas de las nuevas fuentes de datos *60*

2.2.5. Debilidades y retos en el uso de las nuevas fuentes de datos para el estudio de la movilidad urbana *62*

2.3. *Twitter* como fuente de datos para el estudio de la movilidad urbana 68

2.3.1. Introducción a la red social Twitter *68*

2.3.2. Estructura y características de los datos de Twitter *73*

2.3.3. Datos de Twitter vs datos de telefonía móvil *77*

2.3.4. Debilidades de Twitter como fuentes de datos *80*

2.4. Temáticas y aplicaciones de las nuevas fuentes de datos para la investigación de la movilidad metropolitana 83

2.4.1. Aforos y matrices de viajes Origen-Destino *83*

2.4.2. Pautas y recorridos de movilidad individual *85*

2.4.3. Definición de espacios de atracción y generación de viajes *89*

2.4.4. Impactos de eventos en la ciudad *91*

2.4.5. Información para conocer la percepción del transporte *93*

2.5. Aportación de la investigación 95

3. ÁREA DE ESTUDIO, DATOS Y METODOLOGÍA **99**

3.1. Área de estudio 100

3.1.1. El Área Metropolitana de Madrid *100*
3.1.2. La Almendra Central de Madrid *104*
3.1.3. El Metro de Madrid *106*
3.1.4. Las universidades del Área Metropolitana de Madrid *108*
3.1.5. La World Pride 2017 de Madrid *110*
3.2. Datos utilizados 111
3.2.1. Datos de Twitter *111*
3.2.2. Datos de uso del suelo *111*
3.2.3. Datos de población residente, empleo y nivel de renta *112*
3.2.4. Encuesta Domiciliaria de Movilidad 2018 ..... *112*
3.2.5. Datos de número de estudiantes universitarios *113*
3.2.6. Ficheros de transporte privado y público *113*
3.2.7. Datos de uso del Metro de Madrid *113*
3.2.8. Puntos de interés *114*
3.3. Metodología 116
3.3.1. Descarga y almacenamiento de datos *117*
3.3.2. Limpieza, procesado, y enriquecimiento de datos *118*
3.3.3. Agregación y enriquecimiento de los datos. *121*
3.3.4. Análisis y visualización de los datos. *122*
4. CASOS DE ESTUDIO 125
4.1. Diseño y validación de matrices de viajes origen-destino a partir de datos de *Twitter* 126
4.1.1. La movilidad metropolitana a partir de los viajes residencia-trabajo *126*
4.1.2. Metodología específica para el diseño de matrices OD a partir de datos de Twitter *130*
4.1.3. Distribución de los usuarios de Twitter en el espacio según el lugar de residencia y trabajo o estudio *133*
4.1.4. Visualización de matrices OD a partir de datos de Twitter *136*
4.1.5. Calidad y validación de los datos obtenidos *141*
4.2. Visualización de caminos espacio-temporales de movilidad individual a partir de datos de *Twitter* 145
4.2.1. Big Data y la Geografía del Tiempo *145*
4.2.2. Metodología específica para la construcción de caminos espacio-temporales *147*
4.2.3. Distribución de los usuarios de Twitter en el tiempo *151*
4.2.4. Visualización de trayectorias individuales de movilidad en el espacio-tiempo *155*
4.3. Estudio de la movilidad universitaria a partir de datos de *Twitter* 159
4.3.1. Interés de la movilidad universitaria *159*
4.3.2. Metodología específica para el cálculo de áreas de influencia y del modelo de asignación de población universitaria *161*

4.3.3. Lugares de residencia de los usuarios de los campus y áreas de influencia de las universidades *165*

4.3.4. Comparación de tiempos de viajes a partir de transporte público o privado *169*

4.3.5. Modelo de Huff y relación con la asignación de usuarios de Twitter *172*

4.4. Análisis de eventos mediante datos de *Twitter*. El caso de la *World Pride 2017* 176

4.4.1. Los megaeventos y el Big Data *176*

4.4.2. Metodología específica para el análisis del impacto de un evento en la ciudad mediante datos de Twitter *178*

4.4.3. Carácter multicultural del evento. Lugares de procedencia de los visitantes *180*

4.4.4. La impronta espacio-temporal del evento *183*

4.5. Evaluación de las percepciones del Metro de Madrid a través de los textos de *Twitter* 191

4.5.1. La percepción de los sistemas de transporte público *191*

4.5.2. Metodología específica para la extracción de temas y sentimientos de los textos de Twitter *193*

4.5.3. Distribución espacio-temporal de los usuarios de Twitter con sentimiento negativo *199*

4.5.4. Distribución espacio-temporal de los principales temas reportados *202*

4.5.5. Análisis de la causalidad espacial (GWR) *205*

5. CONCLUSIONES 211

5.1. Respuestas a las preguntas de investigación 212

5.2. Conclusiones finales 221

5.3. Futuras líneas de investigación´ 224

5. CONCLUSIONS 227

5.1. Answers to research questions 228

5.2. Final conclusions 236

5.3. Future lines of research 238

PUBLICACIONES, CONGRESOS, Y ESTANCIAS 241

REFERENCIAS BIBLIOGRÁFICAS 243

ÍNDICE DE TABLAS

Tabla 1: Evolución de la ciencia desde la geografía. 47
Tabla 2: Metadatos de un *tweet*. 75
Tabla 3: Aproximaciones a la movilidad a partir de datos de telefonía móvil y de *Twitter*..... 86
Tabla 4: Distribución de la población por zonas metropolitanas. 102
Tabla 5: Estructura de las líneas convencionales de Metro de Madrid. 107
Tabla 6: Universidades del Área Metropolitana de Madrid. 108
Tabla 7: Resumen de fuentes de datos utilizadas en la investigación. 115
Tabla 8: Proceso de filtrado y expansión de datos de *Twitter*. 121
Tabla 9: Resumen de técnicas estadísticas y cartográficas utilizadas en la investigación. 123
Tabla 10: Distribución de la población residencial según zonas metropolitanas. 135
Tabla 11: Distribución de los lugares de trabajo según zonas metropolitanas. 135
Tabla 12: Número de flujos totales de viajes a partir de datos de *Twitter* en las zonas metropolitanas de Madrid..... 140
Tabla 13: Número de flujos porcentuales de viajes a partir de datos de *Twitter* en las zonas metropolitanas de Madrid..... 141
Tabla 14: Flujos de viajes a partir de datos de *Twitter* en las zonas metropolitanas de Madrid clasificados por nivel de residuos..... 143
Tabla 15: Número de usuarios en cada zona de estudio..... 151
Tabla 16: Usuarios de *Twitter* por hora. 152
Tabla 17: Porcentaje de universitarios encontrados en *Twitter* respecto a datos censales por cuartil de renta. 164
Tabla 18: Tiempo medio ponderado a universidades por tipo de transporte..... 170
Tabla 19: Porcentaje de estudiantes asignados por universidad a partir del modelo de *Huff*.... 174
Tabla 20: Palabras claves relacionadas con la World Pride por usuarios de *Twitter*. 179
Tabla 21: Número de usuarios según idioma configurado en *Twitter*. 180
Tabla 22: Frecuencia de *tweets* de cuentas relacionadas con sistemas de transporte público del Área Metropolitana de Madrid durante una semana (16 a 22 de septiembre)..... 194
Tabla 23: Pasos para la limpieza de texto de los *tweets*. 194
Tabla 24: Lista y descripción de los temas formulados. 196
Tabla 25: Parámetros de evaluación del modelo de entrenamiento de sentimiento. 197
Tabla 26: Variables utilizadas en los modelos OLS y GWR. 198
Tabla 27: Número de usuarios de *Twitter* por tema. 203
Tabla 28: Estadísticas de los diagnósticos OLS y GWR..... 206

ÍNDICE DE FIGURAS

Figura 1: Relación entre preguntas de investigación y objetivos específicos de la tesis..... 36
Figura 2: Estructura de la tesis doctoral. 40
Figura 3: Datos producidos en un minuto por diferentes plataformas basadas en las TIC. 45
Figura 4: Las TIC como enlace entre el mundo físico y el virtual. 48
Figura 5: Las escalas de las 3V del Big Data. 49
Figura 6: Dimensiones de los datos geovisualizados. 50
Figura 7: Fuentes de información por frecuencia y valor semántico. 53
Figura 8: Esquema de una red celular de telefonía móvil. 55
Figura 9: Tarjeta inteligente de transporte público de la Comunidad de Madrid. 58
Figura 10: Porcentaje de usuarios de *Twitter* por franja de edad y género. 69
Figura 11: Número de publicaciones por año según la información de Twitter empleada. 73
Figura 12: Estructura de un *tweet*. 76
Figura 13: Esquema de un *datacube*. 87
Figura 14: Mapa del Área Metropolitana de Madrid. 101
Figura 15: Zonas de estudio a analizar dentro del municipio de Madrid. 103
Figura 16: Distritos y barrios de la Almendra Central de Madrid. 105
Figura 17: Red de Metro de Madrid en el Área Metropolitana. 106
Figura 18: Ubicación de los campus universitarios en el Área Metropolitana de Madrid. 109
Figura 19: Pasos de procesamiento del Big Data. 116
Figura 20: Esquema metodológico de la tesis doctoral. 124
Figura 21: Ejemplo de Matriz OD de tiempos de viaje. 128
Figura 22 : Ejemplo de detección de parcela de residencia por moda. 131
Figura 23: Número de usuarios de *Twitter* por lugar de residencia (izquierda) y de trabajo (derecha). 134
Figura 24: Matriz de flujos de viajes a partir de datos de *Twitter* a nivel de municipios y distritos del Área Metropolitana de Madrid. 137
Figura 25: Generaciones y atracciones de Alcobendas y Getafe (a: generaciones de Alcobendas, b: atracciones de Alcobendas, c: generaciones de Getafe, d: atracciones de Getafe). 138
Figura 26: Matriz de flujos de viajes a partir de datos de *Twitter* a nivel de zonas metropolitanas de Madrid. 139
Figura 27: Correlación bivariada entre valor de viajes de *Twitter* y valor de viajes de la EDM a nivel de distritos y municipio (izquierda) y zonas metropolitanas (derecha). 143
Figura 28: Matriz de distribución de residuos de la correlación bivariada entre valores de viajes de *Twitter* y valores de viaje de la EDM (nivel de zonas metropolitanas). 144
Figura 29: Clasificación de fuentes de datos a partir de escalas espacio-temporales 147
Figura 30: Esquema de un camino espacio-temporal. 148
Figura 31: Porcentaje de usuarios de *Twitter* durante el día. 151
Figura 32: Número de usuarios de *Twitter* en diferentes momentos del día. 153
Figura 33: Distribución de usuarios por hora en Puente de Vallecas (1), Nuevos Ministerios-AZCA (2), Ciudad Universitaria (3), y Parque del Retiro (4). 154
Figura 34: Distribución de usuarios por usos de suelo y hora en Puente de Vallecas (1), Nuevos Ministerios-AZCA (2), Ciudad Universitaria (3), y Parque del Retiro (4). 155
Figura 35: Caminos espacio-temporales a lo largo del día (2D) en Puente de Vallecas (1), Nuevos Ministerios-AZCA (2), Ciudad Universitaria (3), y Parque del Retiro (4). 156
Figura 36: Caminos espacio-temporales a lo largo del día (3D). 158

Figura 37: Municipios y distritos del Área Metropolitana de Madrid según nivel de renta por cuartiles. 163
Figura 38: Distribución por campus de los alumnos universitarios a partir de datos oficiales (izquierda) y número de usuarios detectados en *Twitter* (derecha). 165
Figura 39: Número de usuarios residentes detectados en el Área Metropolitana de Madrid a partir de *Twitter*. 166
Figura 40: Porcentaje de población universitaria detectada en *Twitter* por universidad en cada municipio y distrito. 168
Figura 41: Diagrama de caja de tiempos ponderados de viaje a universidad por tipo de transporte. 169
Figura 42: Porcentaje de alumnos residentes por tipo de universidad según grupos de municipios y distritos por nivel de renta. 171
Figura 43: Tiempos de transporte privado y público según cuartiles de municipios y distritos por nivel de renta. 172
Figura 44: Relación entre número de usuarios *Twitter* y estudiantes estimados a partir del modelo de *Huff*. 175
Figura 45: Usuarios detectados en *Twitter* durante la *World Pride* según provincia de procedencia. 182
Figura 46: Usuarios detectados en *Twitter* durante la *World Pride* según país de procedencia. 182
Figura 47: Volumen de actividad por días en la Almendra Central (A), Barrio de Chueca (B), y Aeropuerto (C). 184
Figura 48: Comparación de número de *tweets* por horas en la Almendra Central (A) y en el Barrio de Chueca (B) 185
Figura 49: Número usuarios (izquierda) y cambio porcentual (derecha) respecto a semana habitual en la Almendra Central por barrios durante la *World Pride*. 186
Figura 50: Número usuarios (izquierda) y cambio porcentual (derecha) respecto a semana habitual en la Almendra Central por secciones censales durante la *World Pride*. 187
Figura 51: Clústeres de asociación espacial local en la Almendra Central por secciones censales durante la *World Pride* (izquierda) y durante una semana habitual (derecha). 188
Figura 52: Número de usuarios por secciones censales en la Almendra Central en franjas del fin de semana. 190
Figura 53: Número de usuarios de *Twitter* totales y porcentaje por sentimientos por línea de Metro. 200
Figura 54: Distribución de usuarios de *Twitter* con sentimiento negativo en la red de Metro de Madrid. 200
Figura 55: Relación entre número de usuarios de *Twitter* con sentimiento negativo y número de viajeros registrados en datos oficiales a nivel de estaciones (izquierda) y líneas de Metro (derecha) 201
Figura 56: Porcentaje de usuarios de *Twitter* con sentimiento negativo en el Metro de Madrid por día y hora. 202
Figura 57: Principal tema con sentimiento negativo en las estaciones del Metro de Madrid. 203
Figura 58: Porcentaje de usuarios de *Twitter* con sentimiento negativo por tema y línea. 204
Figura 59: Número de usuarios de *Twitter* con sentimiento negativo en cada tema por día y hora. 205
Figura 60: Distribución residuos y valores R^2 locales en el Área Metropolitana de Madrid. 208
Figura 61: Distribución de los coeficientes de las variables exploratorias utilizadas en el modelo GWR en el Área Metropolitana de Madrid. 209

1. INTRODUCCIÓN

1.1. Interés y oportunidad de la investigación

Actualmente vivimos en una tercera revolución urbana basada en una sociedad globalizada, diversificada e híbrida, con numerosas redes abiertas y especializadas, en la que la población mundial se concentra en los sistemas metropolitanos y las aglomeraciones urbanas (Ascher, 2004). Por primera vez en la historia de la humanidad hay un mayor número de habitantes en las ciudades que en el campo. Como consecuencia, las ciudades van creciendo y expandiéndose. Poniendo como ejemplo Europa, el 75% de su población vive en ciudades.

En este panorama, la movilidad es uno de los grandes retos de las áreas metropolitanas del siglo XXI. La movilidad es entendida como la suma de los desplazamientos individuales de los ciudadanos, con el objetivo de acceder al mercado de trabajo, bienes y servicios (Gasparini & Guidicini, 1990). El crecimiento continuo de las ciudades, la formación y despolarización de áreas metropolitanas y la especialización de zonas (destacando las periferias metropolitanas como áreas residenciales y las áreas centrales como lugares de trabajo, ocio o negocios) conllevan un aumento importante de la demanda de la movilidad y un cambio en sus características, entre los que destacan el aumento del número de viajes, la mayor diversidad de motivos de desplazamiento, la diversificación de horarios, el uso intensivo de transportes motorizados, y trayectos cada vez con mayores distancias de recorrido y mayores tiempos de viaje (Banister, 2011; Gutiérrez & García-Palomares, 2007).

Las nuevas características de la movilidad metropolitana han provocado un aumento de la congestión en los sistemas de transporte, tanto privados como públicos, creando una situación poco sostenible a nivel ambiental (polución atmosférica, contaminación acústica), social (disminución de la accesibilidad en las zonas periféricas o con menor nivel socioeconómico, aumento de la desigualdad social en dichas zonas respecto a áreas centrales) y económico (consumo excesivo de energía y recursos). Ante esta situación, es necesario plantear un modelo de gestión de la movilidad basado en la sostenibilidad ambiental, económica, y social. El objetivo es obtener una dinámica de transporte respetuosa con el medio ambiente, que favorezca el desarrollo económico de las ciudades, y promueva la inclusión social (Gutiérrez-Puebla et al., 2019). Este modelo debe por tanto permitir un consumo eficiente y equilibrado del transporte, que no consuma materia y energía de manera excesiva, que facilite la reducción de la contaminación mediante un

mayor uso de desplazamientos peatonales o en bicicleta y que diseñe una mejor planificación del transporte público (Banister, 2008).

La necesidad de un modelo sostenible para la gestión de la movilidad requiere de datos para comprender el sistema urbano y predecir sus dinámicas tanto a corto como a largo plazo (Cheng, Gould, Han, & Jin, 2016). Para ello hacen falta fuentes de alta resolución tanto espacial como temporal, que permitan obtener información sobre la distribución de la población y sus comportamientos de movilidad en diferentes escalas y casuísticas, y que faciliten la visualización de recorridos de viaje diario con el objetivo de obtener un conocimiento profundo de las dinámicas de la movilidad metropolitana (Rashidi, Abbasi, Maghrebi, Hasan, & Waller, 2017; Schwanen, 2017). La información sobre las pautas de movilidad tiene una importancia vital, pues de ella depende la elaboración de diagnósticos correctos, y la utilización de las herramientas de modelización de las actuaciones a promover y la estimación de sus impactos (Miralles-Guasch, 2012; Miralles-Guasch & Martínez, 2013). En consecuencia, es indispensable contar con datos constantes, de alta resolución espacio-temporal y actualizados que permitan el entendimiento de la movilidad en un sistema metropolitano.

Habitualmente los datos utilizados por los gestores del transporte provienen de encuestas o entrevistas, como las Encuestas Domiciliarias de Movilidad (EDM). La EDM de 2018 organizada por el Consorcio de Transportes de la Comunidad de Madrid combinó en su metodología tanto entrevistas personales como llamadas telefónicas durante cinco meses para cubrir una muestra de más de 85.000 personas[1]. Estas encuestas son de gran interés, pues proporcionan datos bastante detallados y específicos de los viajes y medios de transportes, permitiendo obtener información muy completa sobre la movilidad metropolitana. Sin embargo, este tipo de fuente de datos requiere de enormes muestras, lo cual conlleva un fuerte trabajo de campo, unos tiempos de ejecución altos y unos costes muy elevados de realización. Como consecuencia, los datos de las encuestas presentan una muy baja periodicidad (frecuentemente de diez años; la EDM anterior se publicó en el año 2004) y unos costes económicos de realización bastante elevados. Poniendo como ejemplo de nuevo la EDM de 2018, su coste de realización fue de casi 2.5 millones de euros y su plazo de ejecución de 20 meses[2].

[1] http://www.comunidad.madrid/noticias/2018/02/07/entrevistaremos-85000-personas-mejorar-transporte-region

[2] https://portal-local.es/actualidad-local/economia/item/22007-nueva-encuesta-domiciliaria-para-mejorar-la-planificacion-de-los-servicios-del-transporte-publico-madrileno.html

Ante los inconvenientes de las fuentes tradicionales para obtener datos de forma rápida y constante, es interesante analizar la utilidad de otras fuentes de datos, que proporcionen información de forma rápida y continua, que sea fácil de actualizar, con una escala espacial y temporal adecuada, y con costes económicos menores. En los últimos años ha surgido una serie de herramientas vinculadas a los cambios tecnológicos de las últimas décadas en comunicación y transporte, y al desarrollo de las *Smart Cities*, que permiten obtener nuevos datos que se pueden incorporar a los estudios de movilidad urbana (Miralles-Guasch, Delclòs, & Vich, 2015; Schwanen, 2017). Estas herramientas, las Tecnologías de la Información y Comunicación (TIC) son fuentes de datos basadas en sensores, dispositivos, y la actividad de los usuarios en internet. Entre estas nuevas tecnologías destacan la Web 3.0, los teléfonos móviles inteligentes o *smartphones*, los navegadores GPS, las tecnologías *wireless*, las redes sociales, y las tarjetas inteligentes o *e-cards* (Gutiérrez-Puebla, García-Palomares, & Salas-Olmedo, 2016).

Los ciudadanos anteriormente eran meros receptores pasivos de información, en un flujo hacia abajo en el que la información que se consumía provenía desde las instituciones. Sin embargo, al interactuar con las TIC, los ciudadanos se convierten en productores de cantidades enormes de datos en un flujo hacia arriba en el que las instituciones absorben los datos que genera la población urbana (García-Palomares, Salas-Olmedo, Moya-Gómez, Condeço-Melhorado, & Gutiérrez, 2018; Gutiérrez-Puebla et al., 2016). Vivimos en una sociedad hiperconectada en la que las personas usan y consumen internet de forma masiva y diaria y en la que casi todos los ciudadanos llevan consigo un teléfono móvil que permite la monitorización de su actividad humana diaria. A través de sus dispositivos móviles, los ciudadanos generan una huella digital de sus actividades y movimientos, una traza digital que puede ser seguida (Blanford, Huang, Savelyev, & MacEachren, 2015). Esta traza digital está íntimamente entremezclada con las geografías materiales, offline, de la vida diaria (Jin et al., 2017).

Como consecuencia del uso móvil de internet, y de la penetración extensa del mercado de los dispositivos móviles en muchos países, en solo unas décadas se ha pasado de una escasez de datos disponibles, a otra de sobreabundancia de datos de distinto tipo que son generados de forma diaria. Se crea un flujo enorme de datos que se recogen de forma pasiva sin que el usuario tenga que hacer nada, y se distribuyen y hacen accesible a cualquier lector desde cualquier punto del planeta (Gutiérrez-Puebla et al., 2016; Kitchin, 2013). La habilidad actual para adquirir, procesar, compartir y analizar grandes

cantidades de datos no tiene precedentes en la historia. Se calcula que en los últimos tres años se han generado tantos datos como en toda la historia de la humanidad (Gutiérrez-Puebla et al., 2019).

No es de extrañar que en los últimos años se haya vuelto popular el término *Big Data*, con el que se describen los datos generados a partir de TIC, datos obtenidos en cantidades enormes, a una alta velocidad y desde una gran variedad de fuentes. Estos datos tienen amplias ventajas, que las fuentes de datos tradicionales no pueden ofrecer, como su gran tamaño, su alto dinamismo, su alto detalle espacial y temporal, su bajo coste de preparación, su gran frecuencia y su fácil actualización a tiempo real. El *Big Data* se puede ver como un gran repositorio con datos de la actividad de la gente y de cómo unos conectan con otros (Miller, 2010). El auge del *Big Data* ha permitido la transición de una época donde uno de los retos de la investigación era la obtención de datos debido a su escasez a otra en la que el reto principal en la investigación consiste en la capacidad de seleccionar y procesar adecuadamente los datos necesarios (Gutiérrez-Puebla et al., 2016).

Cada vez hay un mayor número de instituciones y empresas que basan sus estudios en el *Big Data* y usan estos datos masivos para llevar a cabo procesos de control y gestión (Batty, 2013; M. Chen, Mao, & Liu, 2014). El *Big Data* tiene gran valor para las empresas para analizar el comportamiento de los consumidores, diseñar estrategias de geomarketing, o predecir las tendencias del mercado. Por otra parte, las empresas venden estos datos o estudios basados en ellos. Compañías tecnológicas como *Google* o *Facebook* tienen una fuerte capitalización bursátil gracias al gran valor económico de los datos que generan sus usuarios (Gutiérrez Puebla, 2018).

Se estima que el 80% de los datos del *Big Data* están geolocalizados, al contar con coordenadas de latitud y longitud, o por el propio contenido de los datos (Leszczynski & Crampton, 2016). Gracias a la componente espacial de los datos y a su compatibilidad con los Sistemas de Información Geográfica (SIG), es posible realizar análisis geoestadísticos y cartografiar la información procedente de las nuevas fuentes de datos. La combinación de datos de las TIC y de paquetes de software como los SIG para analizar y visualizar los datos se han vuelto indispensables para la investigación geográfica (Ash, Kitchin, & Leszczynski, 2018). Se están abriendo nuevas posibilidades y líneas de estudio en las ciencias sociales y la geografía gracias a la posibilidad de cartografiar y promover el valor de los datos obtenidos por las TIC a una mayor audiencia (Kitchin, 2013). Se

puede decir por tanto, que el futuro de la geografía y el *Big Data* va de la mano (Graham & Shelton, 2013).

Las tecnologías digitales están teniendo efectos cada vez más profundos en la gestión del territorio, la movilidad, y la promulgación de políticas de conocimiento espacial (Ash et al., 2018). Las nuevas fuentes de datos permiten el análisis de pautas espaciales y multitud de procesos que no pueden ser estudiados con estadísticas oficiales o encuestas, con lo que surge la oportunidad de estudiar temas que no habían sido tratados o que habían sido dejados de lado ante la falta de información que ofrecían las fuentes tradicionales (Gutiérrez-Puebla et al., 2016). En el campo de la movilidad, el *Big Data* constituye una materia prima muy valiosa para el estudio del comportamiento urbano, ya que, al conocer la localización cambiante de cada persona a partir de su huella digital, se pueden analizar sus pautas generales de movilidad en el espacio y el tiempo.

Al generarse constantemente grandes cantidades de datos en tiempo real y de forma masiva, es posible tener una nueva visión de la movilidad en nuestras ciudades a partir de datos hasta hace poco tiempo no disponibles, ya que los ordenadores no eran lo suficientemente potentes, los datos no tenían un alto nivel de detalle o no eran de libre acceso, y no existían tecnologías como *Twitter* o los teléfonos inteligentes (Cheshire & Uberti, 2016). Con las nuevas fuentes de datos se pueden captar datos sobre la movilidad urbana o metropolitana, mejorar nuestros estudios de accesibilidad, o estudiar la movilidad en días o momentos específicos. Así, a partir del *Big Data* podemos definir espacios de generación y atracción de viajes, obtener matrices origen-destino, visualizar el recorrido habitual de una serie de individuos, o cartografiar los principales problemas que reportan los usuarios de un servicio de transporte.

Las fuentes de datos basadas en las TIC que se pueden emplear para el estudio de la movilidad urbana presentan diferentes características entre ellas. Las redes sociales son una de las nuevas fuentes de datos más atractivas para la investigación debido a que se han convertido en un componente integral de las sociedades modernas. Se estima que un tercio de la población usa una o más redes sociales (de Smith, Goodchild, & Longley, 2018). Numerosos usuarios publican momentos de su vida, suben fotografías, o comparten opiniones en páginas como *Facebook* o *Instagram*. Muchas empresas usan además las redes sociales como recursos para obtener información, estudiar campos de actuación, o publicitarse. La combinación de la tecnología móvil con las redes sociales permite tener un alto detalle espacial y temporal de momentos de la vida diaria de las

personas, permitiendo la ejecución de análisis de gran interés para entender la movilidad tanto general como de distintos grupos sociales. Sin embargo, este desarrollo conlleva el surgimiento de cuestiones éticas de seguridad y privacidad que hay que tener en cuenta a la hora de realizar análisis de la movilidad humana. Además, redes sociales muy populares y con mucha actividad como *Facebook* o *Instagram* no comparten los datos que obtienen ya que cuentan con alto valor económico que explotan para análisis internos.

Twitter es una de las redes sociales con mayor penetración en todo el mundo debido a su sencillez, fácil uso, y gran compatibilidad con cualquier teléfono móvil (Murthy, 2018). Un usuario puede publicar un *tweet* de forma sencilla, casi instantánea y desde cualquier parte del mundo. Los *tweets* presentan una estampa temporal que indica la fecha completa en la que fueron publicados, y gracias a la tecnología GPS del teléfono móvil, pueden poseer también datos espaciales de latitud y longitud. Además, es posible descargar de forma gratuita muestras de datos geolocalizados de *Twitter*, lo que posibilita realizar estudios de las distribuciones cambiantes de la población en el tiempo (Ciuccarelli, Lupi, & Simeone, 2014; García-Palomares et al., 2018; Longley & Adnan, 2016). Disponemos, por tanto, de una nueva fuente de datos de alto detalle espacio-temporal, gratuita, y de descarga rápida, que permite una actualización constante de los datos, y le confiere de un gran potencial para el análisis de la movilidad metropolitana a tiempo casi real. Los *tweets*, al ser tratados con forma de punto, pueden ser incorporados fácilmente en un SIG, y por tanto analizados y cartografiados (Blanford et al., 2015; Goodchild, 2007).

Sin embargo, los datos de la mayoría de las fuentes asociadas al *Big Data*, incluyendo las redes sociales como *Twitter*, no son creados con el fin de analizar la movilidad urbana y presentan en consecuencia limitaciones o condicionantes a considerar. Esta situación es común en el uso de redes sociales en disciplinas como el planteamiento urbano y la movilidad. En consecuencia, el uso de los datos de las TIC para estudios de movilidad posee debilidades importantes que hay que tener en cuenta en comparación con los datos de las fuentes creadas específicamente para este propósito, como las encuestas de movilidad. Uno de los factores condicionantes es la calidad y propósito de los datos y su posible falta de estructura (Lansley, Smith, Goodchild, & Longley, 2018; Miller & Goodchild, 2014). Los procesos de limpieza y preprocesado de los datos son muy importantes para que las bases de datos solamente contengan datos depurados de errores con los cuales se pueda obtener información confiable. Además, son también importantes el diseño de metodologías y algoritmos para el procesado y análisis estadístico de los

datos, ya que no existe una única metodología estandarizada o consensuada para tratar los datos, sino líneas generales de análisis que pueden dar distintos resultados según la metodología a usar.

Otro factor a tener en cuenta es la limitación de los datos. La resolución temporal de los datos de las redes sociales depende de la frecuencia con la que un usuario genera mensajes, dificultando, por ejemplo, la obtención de datos de sitios de actividades en lugares que no sean de residencia o trabajo, y obligando a que los datos sean recopilados durante mayores periodos temporales. Además, los datos geolocalizados pueden estar sesgados, ya que los usuarios de las nuevas tecnologías como redes sociales suelen tener un perfil sociodemográfico determinado (principalmente población joven de 20-49 años con estudios universitarios) (Gutiérrez Puebla, 2018). Finalmente, el uso de estas nuevas fuentes de datos requiere de un proceso de validación de resultados con el que transformar los datos en información y conocimiento. La solución pasa por emplear metodologías que aprovechen al máximo las fortalezas de las nuevas fuentes de datos e integrarlas con datos de otras fuentes, tanto tradicionales como nuevas, para obtener información de calidad y actualizada sobre la movilidad. No se trata de usar las TIC para sustituir completamente a las fuentes tradicionales, sino aunar y usar conjuntamente las fortalezas de cada tipo de datos.

Esta tesis doctoral busca analizar el valor que tienen las TIC en general, y la red social *Twitter* en particular, como fuentes de datos para el estudio de movilidad. Para ello, la tesis plantea investigar en profundidad la utilidad de los datos que posee un *tweet* (coordenadas espaciales, registros temporales, texto, etc.) y aplicarlos en el estudio de diferentes aspectos de la movilidad metropolitana. En concreto se ha utilizado *Twitter* para obtener matrices Origen-Destino de viajes residencia a trabajo, para la visualización de caminos espacio-temporales individuales, para el estudio de la movilidad universitaria, para observar las pautas de distribución de la población durante eventos, o para el análisis de los comentarios que los usuarios realizan sobre un sistema público de transporte. El estudio de la utilidad de las nuevas fuentes de datos como *Twitter* se ha empezado a estudiar recientemente en esta década y sus bases se están asentando actualmente (Batty, 2013; Schwanen, 2017). Aunque hay trabajos previos que utilizan estas nuevas fuentes en el estudio de la movilidad, se trata de un campo de trabajo que todavía está en desarrollo, con contribuciones y aportes a diferentes situaciones de movilidad en el marco urbano que todavía están siendo estudiadas y analizadas. A nivel nacional son escasos los

trabajos que tratan este tipo de fuentes, y el estudio de este tipo de investigaciones apenas ha sido aplicado en profundidad al marco geográfico de un área metropolitana española como Madrid.

1.2. Preguntas de investigación

En el apartado anterior se ha presentado el interés de esta tesis tanto para la investigación científica geográfica como para la explotación por parte de empresas y organismos. A continuación, se elaboran las preguntas de investigación que este trabajo tratará de responder, y que servirán de guía para el desarrollo de los objetivos de esta investigación:

a) ¿Las fuentes de datos generadas por las TIC son adecuadas para el estudio de la movilidad urbana?

Esta primera pregunta busca entender la validez y adecuación de las nuevas fuentes de datos basadas en las TIC para solucionar problemas relacionados con el transporte o la movilidad urbana. Para ello, es de gran interés evaluar y comparar de manera teórica las ventajas y desventajas de estas nuevas fuentes de datos respecto a las fuentes tradicionales, como las encuestas, y analizar la efectividad de ambas fuentes de datos en el campo de los estudios urbanos en general, y de la movilidad urbana en particular.

b) ¿Qué aporta cada una de las nuevas fuentes de datos al estudio de la movilidad urbana?

Con esta pregunta se pretende clasificar, evaluar y poner en valor las distintas fuentes de datos basadas en las TIC, analizando las ventajas y desventajas de cada una de ellas, con el objetivo de seleccionar una fuente de datos que sea de fácil acceso, y que permita obtener información de los patrones de movilidad de los ciudadanos con el mayor detalle espacial y temporal posible. Dentro de esta pregunta, es de interés comparar y clasificar las distintas redes sociales, y observar de nuevo cuales permiten un fácil acceso y descarga de los datos.

c) ¿Qué herramientas y técnicas pueden ayudar a convertir los datos de redes sociales (como Twitter) en información y conocimiento sobre movilidad?

Aunque aúnan importantes ventajas para su aplicación, los datos procedentes de redes sociales como *Twitter* cuentan con una serie de retos que se están investigando actualmente. Uno de los desafíos más importantes es el estudio de la metodología

adecuada para la descarga, preparación, limpieza y procesado de los datos. Otro componente por analizar son las técnicas geoestadísticas, cartográficas y de visualización que permitan el análisis de los datos y la obtención de resultados, es decir, información que se pueda convertir en conocimiento.

d) ¿Pueden los datos de Twitter ser usados para obtener matrices de viajes en espacios metropolitanos?

Los datos de *Twitter* pueden estar georreferenciados o geolocalizados, por lo que, dependiendo del lugar y el momento, es posible estimar lugares de residencia o trabajo. Esta pregunta busca responder si el alto granulado espacial que se le supone a los datos de *Twitter* permite analizar, visualizar, y comprender los fenómenos ubicados en el espacio como los viajes residencia-trabajo. Se busca investigar la precisión espacial de los *tweets* para la posible identificación de lugares de residencia o trabajo, como influye la escala espacial en los datos, y si los resultados obtenidos se ajustan a la realidad.

e) ¿Es posible utilizar los datos de Twitter para visualizar la movilidad metropolitana mediante caminos espacio-temporales?

Los datos basados en *Twitter* no cuentan solamente con coordenadas espaciales, sino también registros temporales que dan información sobre la hora y fecha en la que fue publicado un *tweet*. Esta pregunta investiga la fiabilidad del alto granulado temporal que también se les otorgan a estos datos para poder estudiar si los distintos usos del suelo que los ciudadanos utilizan a lo largo del día se ajustan a la realidad, y la capacidad de *Twitter* para representar los trayectos espacio-temporales que realizan los individuos de forma regular a partir de sus coordenadas tanto espaciales como temporales.

f) ¿Se puede estudiar con los datos de Twitter la movilidad de población vinculada a espacios concretos de la ciudad?

Pese a que las nuevas fuentes de datos basadas en las TIC cuentan con un alto detalle espacio-temporal, estos datos suelen carecer de datos temáticos, sociales, o económicos, haciendo necesario el enriquecimiento de los datos en determinados estudios. Partiendo de esta desventaja, se busca investigar la capacidad de *Twitter* para conseguir información de los comportamientos de movilidad asociada a determinados espacios que atraen una población con perfiles sociodemográficos específicos, a partir de la identificación de posibles miembros en base a su huella digital generada.

g) ¿Son los datos de Twitter útiles para analizar el impacto de eventos en el comportamiento espacial de la población?

La organización y celebración de un evento o un festival importante conlleva un escenario con unos patrones espacio-temporales específicos y unos comportamientos de uso de la ciudad que difieren del uso habitual en días regulares. A partir de ahí, se busca averiguar la efectividad de los datos de *Twitter* para monitorizar tanto desde el punto de vista espacial como el temporal la distribución de la población en un área de estudio durante un festival, observar diferencias respecto a un escenario habitual, e identificar el lugar de procedencia de la población visitante.

h) ¿Puede la información semántica de los textos de Twitter ser válida para el estudio del transporte urbano?

Una de las debilidades de las nuevas fuentes de datos en comparación a las fuentes tradicionales como las encuestas radica en el valor cualitativo de los datos. Los datos descargados de *Twitter* tienen un cierto nivel de riqueza semántica al contar con un campo de texto donde los usuarios escriben sus opiniones o sentimientos acerca de un tema, por lo que con técnicas de minería de texto se puede extraer información complementaria. Con esta pregunta, se plantea analizar el valor semántico de *Twitter* a partir de la extracción de información específica como temas y sentimientos, y poner su puesta en valor en el campo del transporte urbano.

i) ¿Hasta qué punto son los datos de Twitter válidos para el estudio de la movilidad en los espacios metropolitanos?

Esta última pregunta busca valorar y sintetizar desde una perspectiva global todos resultados, respuestas y conclusiones obtenidos para responder a las anteriores preguntas, con el propósito de formular una respuesta que permita cumplir con el objetivo principal de la tesis. Se busca, por tanto, resolver y validar si las nuevas fuentes de datos pueden elaborar una simulación fiel a la realidad, y si los resultados obtenidos son similares a la información proporcionada por fuentes de datos oficiales como las EDM.

1.3. Objetivos

La tesis doctoral parte de la hipótesis de que el desarrollo de los sistemas de telefonía y navegación móvil y las plataformas de internet han propiciado la creación de un espejo a partir de las huellas digitales generadas por los ciudadanos que emplean estas tecnologías en su vida diaria. Por tanto, es posible acceder a una simulación que refleja los comportamientos metropolitanos de movilidad y usos del suelo, permitiendo obtener datos de una forma mucho más rápida, constante y económica que mediante métodos convencionales.

Por tanto, el **objetivo general** de investigación de la tesis es estudiar el valor de las nuevas fuentes de datos originadas a partir de las TIC como métodos alternativos a las fuentes de datos originales para el análisis de la movilidad metropolitana, y analizar el potencial de estos datos a partir de sus diferentes ángulos (espacial, temporal, y semántico) para elaborar análisis y diagnósticos de movilidad con el fin de lograr soluciones alternativas y sostenibles para la gestión, planificación y crecimiento de las áreas metropolitanas en diversos escenarios.

Dentro de este objetivo general, la red social *Twitter* cuenta con un valor añadido al ser una fuente de datos de fácil acceso y que permite obtener en poco tiempo datos que pueden ser analizados y cartografiados rápidamente mediante el uso de los SIG. Por tanto, es de interés investigar la relevancia particular de esta plataforma para analizar patrones espaciales, temporales y temáticos, e identificar posibles respuestas y soluciones que contribuyan a un entendimiento profundo de la movilidad en un área metropolitana.

Si bien cada vez van surgiendo más estudios de análisis urbanos mediante el uso de nuevas fuentes de datos, sigue siendo un campo con un bajo número de investigaciones en el ámbito europeo en general, y en España en particular. Por ello, esta tesis quiere aprovechar el uso de las nuevas fuentes de datos para analizar y visualizar en profundidad los comportamientos urbanos en el Área Metropolitana de Madrid, una de las zonas con mayor población y valor económico del continente europeo, y que sin embargo cuenta con pocos estudios al respecto.

En esta tesis de investigación, se marcan una serie de objetivos específicos a partir de las preguntas de investigación, la hipótesis y el objetivo principal desarrollados en los párrafos anteriores. La figura 1 al final del apartado marca la relación entre las preguntas

de investigación, los objetivos específicos, y los capítulos de la tesis, formulando el armazón de esta investigación científica.

1. Realizar una revisión bibliográfica y desarrollar un marco teórico que permita comprender el papel de las nuevas tecnologías como fuentes de datos para el estudio de la movilidad, las fuentes disponibles, las temáticas abordadas, y las metodologías utilizadas hasta ahora.

Las nuevas fuentes de datos basadas en el *Big Data* son muy variadas y aunque cuentan con rasgos generales como las 3 "V" (velocidad, variedad, volumen), se diferencian en dos aspectos de interés: el valor semántico que poseen, y la frecuencia con la que generan datos, afectando al granulado espacio-temporal de los mismos. Por tanto, es importante comprender las características de las TIC y su aportación en diferentes estudios de movilidad urbana.

En toda investigación, la búsqueda, recopilación, lectura y síntesis de investigaciones anteriores o en curso que aborden la misma temática es crucial. De este modo es posible realizar un marco teórico sólido y sustentarse en las metodologías de otros trabajos para poder desarrollar una metodología propia y original. Esta revisión bibliográfica se hace constantemente a lo largo de la investigación, para poder facilitar el desarrollo de un marco teórico actualizado y con rigurosidad científica.

Este marco teórico es la guía para comprender y poner en valor, teórico y práctico, las nuevas fuentes de datos basadas en el *Big Data*, comprender sus ventajas y analizar sus debilidades y desafíos de cara al futuro. Igualmente, se busca clasificar las distintas fuentes de datos basadas en las TIC, y ordenar las investigaciones anteriores para facilitar la investigación propia desde diferentes vertientes y oportunidades.

2. Investigar el valor de Twitter como fuente de datos alternativa a las fuentes tradicionales para el estudio de la movilidad, comprender su funcionamiento y la estructura de sus datos.

Este objetivo parte como consecuencia de resolver el primer objetivo específico. *Twitter* es una de las fuentes basadas en las nuevas tecnologías más empleadas en los últimos años por investigadores en estudios urbanos. Con este objetivo se busca analizar en profundidad las características de esta red social, tanto en volumen, variedad y velocidad, como en su detalle espacio-temporal y su valor semántico.

Para ello es necesario entender la terminología relacionada con esta red social (*trending topics*, *hashtags*, *followers*, etc.) y comprender la estructura detrás de un *tweet* y los metadatos de interés para la investigación de temas urbanos. De nuevo, es necesario elaborar un marco teórico sustentado en bibliografía específica que permita aprender diversas metodologías que emplean datos de *Twitter* en análisis urbanos, y entender las ventajas y debilidades del uso de esta red social mediante la comparación con otras nuevas fuentes utilizadas en el estudio de la movilidad, como por ejemplo los datos de telefonía móvil.

3. Proponer y establecer una metodología adecuada para la descarga, almacenamiento, procesado y el enriquecimiento de datos basados en Twitter con el fin de utilizarlos para el estudio de la movilidad urbana.

Un inconveniente de los datos descargados de *Twitter* radica en la propia naturaleza de los *tweets*: no son datos diseñados con el objetivo de estudiar patrones de movilidad urbana o de usos del suelo. Sin embargo, al tener datos geolocalizados y con registro temporal, es posible convertir estos datos en información útil para el análisis de la movilidad. Para ello, es necesario establecer una metodología que permita transformar estos datos y hacerlos viables para el estudio de la movilidad urbana.

Partiendo de los datos brutos, es de interés establecer un método de descarga de forma que permita una actualización rápida y constante. Una vez obtenidos los datos, se evalúan métodos de almacenamiento eficaces y que faciliten la integración de los *tweets* en bases de datos espaciales que puedan tratarse en SIG. A continuación, hay que limpiar los datos y eliminar cuentas *bot* o robot que no representan usuarios válidos, y tener en cuenta que usuarios no son de interés para la investigación (usuarios con poca movilidad o con poco número de *tweets*). Finalmente, se busca dotar a los *tweets* de contenido adicional como información de unidades espaciales, datos de población, uso del suelo, u otro tipo de datos temáticos de interés, o separar la información temporal para posteriormente poder agrupar *tweets* en diferentes escalas temporales.

4. Evaluar la capacidad de los datos de Twitter para el análisis espacial de la movilidad metropolitana y diseñar una metodología de visualización de flujos de viajes a partir del diseño de matrices Origen-Destino.

El mayor punto de interés en la investigación de esta tesis radica en el uso de datos con coordenadas georreferenciadas para el análisis de la movilidad metropolitana. Por tanto,

la dimensión espacial de los datos cobra una importancia clave. Los principios de la geografía como la primera ley de *Tobler*, o principio de autocorrelación espacial, se deben cumplir en los patrones espaciales obtenidos a partir de datos de *Twitter*. Es importante analizar el detalle espacial de los datos descargados usando otras fuentes de datos para poder ver si la simulación se parece espacialmente a la realidad que reflejan las fuentes oficiales.

Las matrices Origen-Destino son una de las herramientas más usadas para visualizar relaciones espaciales o flujos de movilidad entre distintas unidades espaciales. Teniendo en cuenta que cada municipio o distrito de Madrid es una unidad espacial, uno de los objetivos de este trabajo es el uso de datos de *Twitter* para evaluar su capacidad de mostrar patrones de distribución espacial en un área de estudio, observar cuáles son las principales zonas de residencia y trabajo de la población, y diseñar matrices Origen-Destino con las que poder visualizar las relaciones entre las unidades espaciales que componen el área metropolitana a diferentes escalas.

5. Utilizar los datos de Twitter para visualizar trayectorias de movilidad individual que incluyan el componente temporal mediante la construcción de caminos espacio-temporales.

Una persona no solo está ubicada en un punto en el espacio, sino también en el tiempo. En análisis urbanos no se puede entender la dimensión espacial sin la temporal y viceversa. Por tanto, este objetivo marca la investigación de los datos de *Twitter* como fuentes válidas para estudiar la dimensión temporal de la movilidad metropolitana, aprovechando que cada dato cuenta con un registro temporal completo de fecha y hora. Al contrario que los resultados obtenidos por análisis espaciales, es más difícil visualizar en un SIG patrones temporales. Sin embargo, el desarrollo de la geografía del tiempo y de los propios SIG han permitido que en esta última década se estén empezando a realizar análisis y visualizaciones temporales a partir de bases de datos geográficas.

Este desafío también se marca en el actual objetivo, que busca el empleo de caminos espacio-temporales 3D como herramienta de visualización de los movimientos de las personas en el espacio a lo largo del tiempo. Además, se plantea la monitorización de cambios en los usos del suelo a lo largo del día en determinadas zonas de usos con perfiles de actividad específica, para poder visualizar si los procesos temporales observados a partir de *Twitter* se ajustan con la realidad.

6. Analizar patrones de movilidad de grupos vinculados a espacios de atracción como los campus universitarios utilizando datos de Twitter.

Un colectivo social suele estar definido por determinadas características socioeconómicas y demográficas. Generalmente, estas características también afectan directamente a los hábitos de residencia o movilidad, por lo que pueden generar huellas digitales con patrones particulares en comparación a los registros generados por otros individuos, como la utilización de determinados espacios a horas concretas del día. Sin embargo, para poder identificar a un colectivo social, es necesario disponer de una información temática (nivel de renta, edad, género, etc.) que las nuevas fuentes de datos como *Twitter* no disponen.

Ante esta situación, un objetivo del trabajo radica en estudiar los patrones de movilidad particulares de espacios determinados que sirven como zonas de atracción de viajes para un colectivo concreto, como es el caso de la población universitaria. Para ello se busca identificar a esta población mediante métodos como la selección de *tweets* ubicados en estos espacios (en este caso campus universitarios). Aprovechando esa selección espacial, los sesgos de las propias redes sociales, y el enriquecimiento de datos con otras fuentes oficiales, es posible asumir que se está trabajando con la población que trabaja o estudia en una universidad, y extraer sus patrones particulares de movilidad.

7. Evaluar la funcionalidad de los datos de Twitter para la identificación de visitantes y el análisis del impacto espacio-temporal de un evento en la ciudad.

Las ciudades, aunque heterogéneas, suelen presentar rasgos de movilidad y usos del suelo dentro de unos patrones regulares a lo largo del año. Sin embargo, la organización y celebración de un evento o festival altera estos patrones de movilidad y distribución y de la población, entre otras causas por el elevado número de visitantes que viaja a la ciudad en las fechas en las que se celebra el evento.

Por tanto, el siguiente objetivo busca usar los datos de *Twitter* para comparar el ritmo espacio-temporal de la ciudad durante un evento respecto a los patrones visualizados durante un periodo de actividad habitual, y observar las diferencias que se producen como el aumento del uso de determinados espacios. Además, no solo se busca estudiar *Twitter* como una fuente alternativa para el estudio del cambio morfológico urbano durante un evento, sino que también se plantea desarrollar una metodología que permita la identificación y conteo de visitantes, y su provincia o país de origen.

8. Analizar los sentimientos de los usuarios de un sistema de transporte público a partir de los textos de Twitter mediante identificación de temas y análisis semántico.

El campo principal de un *tweet* es su texto. Al igual que con los atributos espacial y temporal de los datos podemos responder a las preguntas "donde" y "cuando", el texto de los *tweets* puede servir para responder a la pregunta "que" correspondiente con la tercera dimensión de los datos, la dimensión semántica. Aun así, el carácter desestructurado de los *tweets* y su brevedad semántica al tratarse de textos con un número de caracteres limitados, dificulta la extracción de información semántica de utilidad para el planteamiento urbano.

Con este objetivo, se plantea un análisis semántico de los textos de *Twitter* para extraer temas relacionados con el uso de un sistema de transporte público, ya que estos medios de locomoción son la cara visible de la movilidad diaria para los ciudadanos. A partir de una metodología de agrupación de los textos de los *tweets* en temas y sentimientos, se busca ubicar distintas percepciones y problemas en el espacio que ocupa una red de transporte público e identificar posibles puntos de actuación en el que puedan intervenir los gestores del transporte para mejorar el sistema.

9. Explorar los resultados obtenidos a partir de los datos de Twitter y poner en valor el uso de las nuevas fuentes de datos basadas en las TIC frente a las fuentes tradicionales de datos.

Como se ha comentado previamente, la hipótesis de esta investigación radica en que la huella generada en el mundo digital es un reflejo tanto espacial como temporal de la actividad desarrollada en el mundo real. Por tanto, los resultados obtenidos a partir de fuentes de datos como *Twitter* deberían tener un alto grado de similitud con la información obtenida a partir de datos oficiales como los censos de población.

A partir de la comparación de los resultados obtenidos por los datos de *Twitter* frente a datos de fuentes oficiales, este objetivo busca validar el resto de objetivos específicos marcados y permitir responder al objetivo principal de la tesis. Además, al evaluar la eficacia de *Twitter* como fuente de datos alternativa, se investiga también la eficacia de las fuentes de datos tradicionales para el estudio de la movilidad metropolitana, con el propósito de entender las fortalezas y debilidades de ambas fuentes de datos para poder así complementarlas de la forma más eficaz posible en futuras líneas de investigación.

Figura 1: Relación entre preguntas de investigación y objetivos específicos de la tesis.

Fuente: Elaboración propia.

1.4. Estructura del trabajo

Esta tesis está estructurada en cinco capítulos. La Figura 2 recoge y resume la estructura de la tesis. El primer capítulo introduce la investigación y hace énfasis en la relevancia del interés del tema tanto en el ámbito de la investigación como en el empresarial. Se formula el objetivo principal y una serie de objetivos específicos diseñados para responder a una serie de preguntas de investigación previamente descritas. Además, se describe la estructura de la tesis y la vinculación de cada una de sus partes con los objetivos marcados.

El marco teórico que sustenta esta tesis doctoral lo conforma el segundo capítulo de la tesis. Este capítulo introduce el papel y las características de las nuevas fuentes de datos asociadas al *Big Data* en el estudio de la movilidad. Primero se introducen las fuentes tradicionales de datos y se analizan sus ventajas e inconvenientes, para enseguida presentar y definir las nuevas fuentes de datos. A continuación, se estudian sus principales características, se formula una clasificación basada en el tipo de fuentes, y se contemplan las ventajas y debilidades de su uso en los análisis de movilidad. Después, se entra en mayor profundidad a la herramienta principal utilizada para todas las investigaciones de esta tesis: la red social *Twitter*. Se analiza la estructura y funcionamiento de la red social, sus mensajes y su API, para poder comprender en detalle el interés de esta fuente de datos respecto a otras fuentes basadas en las TIC, y para poder enseñar que metadatos son los más relevantes para la investigación realizada. Además, se comparan las características principales de esta red social respecto a los datos de telefonía móvil, y se analizan los sesgos e inconvenientes de esta *Twitter* como fuente de datos, para establecer una serie de desafíos. Por último, se pone en contexto las oportunidades del uso de las fuentes de datos basadas en las TIC en diversas investigaciones en el estudio de la movilidad urbana. Estos trabajos se clasifican por temática y aplicaciones. Una vez definido el estado del arte, se establece la aportación científica de esta tesis doctoral. El marco teórico formado por este capítulo corresponde con el artículo científico titulado "*Nuevas fuentes y retos para el estudio de la movilidad urbana*", publicado en la revista *Cuadernos Geográficos* en el año 2017. Para esta tesis se ha realizado una nueva revisión del estado del arte, conllevando una redacción actualizada.

El tercer capítulo introduce el área de estudio de la investigación, en este caso el Área Metropolitana de Madrid. Se provee un contexto local para entender las características de este espacio metropolitano y su interés para la investigación conducida, y se presentan

también las diferentes infraestructuras del área de estudio que se estudiarán en los casos de estudio que conforman el siguiente capítulo. A continuación, se enumeran los datos empleados durante la investigación y se identifican sus fuentes. A su vez, se establece una metodología de descarga, almacenamiento, procesado y enriquecimiento de los datos de *Twitter* para su posterior uso en los casos de estudio de la tesis, y, se resumen las técnicas de análisis y visualización de datos empleadas en los posteriores casos de estudio.

El cuarto capítulo desarrolla una serie de casos de estudio que muestran el uso práctico de *Twitter* como fuente de datos para el estudio de la movilidad urbana. El primer caso de estudio establece una metodología para identificar potenciales lugares de residencia y trabajo y obtener a partir de ellos matrices Origen-Destino de viajes a distintas escalas. Las matrices obtenidas se han validado mediante el uso de fuentes de datos oficiales como las encuestas domiciliarias de movilidad. Esta sección se corresponde con la investigación titulada "*Social media and urban mobility: Using Twitter to calculate home-work travel matrices*", publicada en la revista *Cities* en el año 2019.

El segundo caso de estudio representa la trayectoria regular de los usuarios de *Twitter* de la muestra que se ubican en varias zonas del área de estudio, cada una de ellas con un perfil específico de actividades y usos del suelo. Para ello, se desarrolla un método de visualización 3D basado en caminos espacio-temporales que permita representar la movilidad del individuo no solo en el espacio, sino también en el tiempo. Además, también se representa el uso del suelo principal en cada punto espacio-temporal del camino. Este capítulo se corresponde con el trabajo titulado "*Spatio-temporal mobility and Twitter: 3D visualization of mobility flows*", enviado a la revista *Journal of Maps* en el año 2020.

El tercer caso de estudio propone un método de identificación de población universitaria y del campus al que asisten estos usuarios a partir de datos de *Twitter*, y analiza los patrones de movilidad residencia-universidad mediante parámetros como el tipo de universidad, el nivel de renta del municipio o distrito de origen, o los tiempos de viaje al campus según el modo de transporte. Además, se establece un modelo gravitacional de *Huff* para validar la eficacia de los datos de *Twitter*. Este capítulo se corresponde con el artículo titulado "*Big Data y universidades: análisis de movilidad de los estudiantes universitarios a partir de datos de Twitter*", publicado en la revista *Geofocus* en el año 2019.

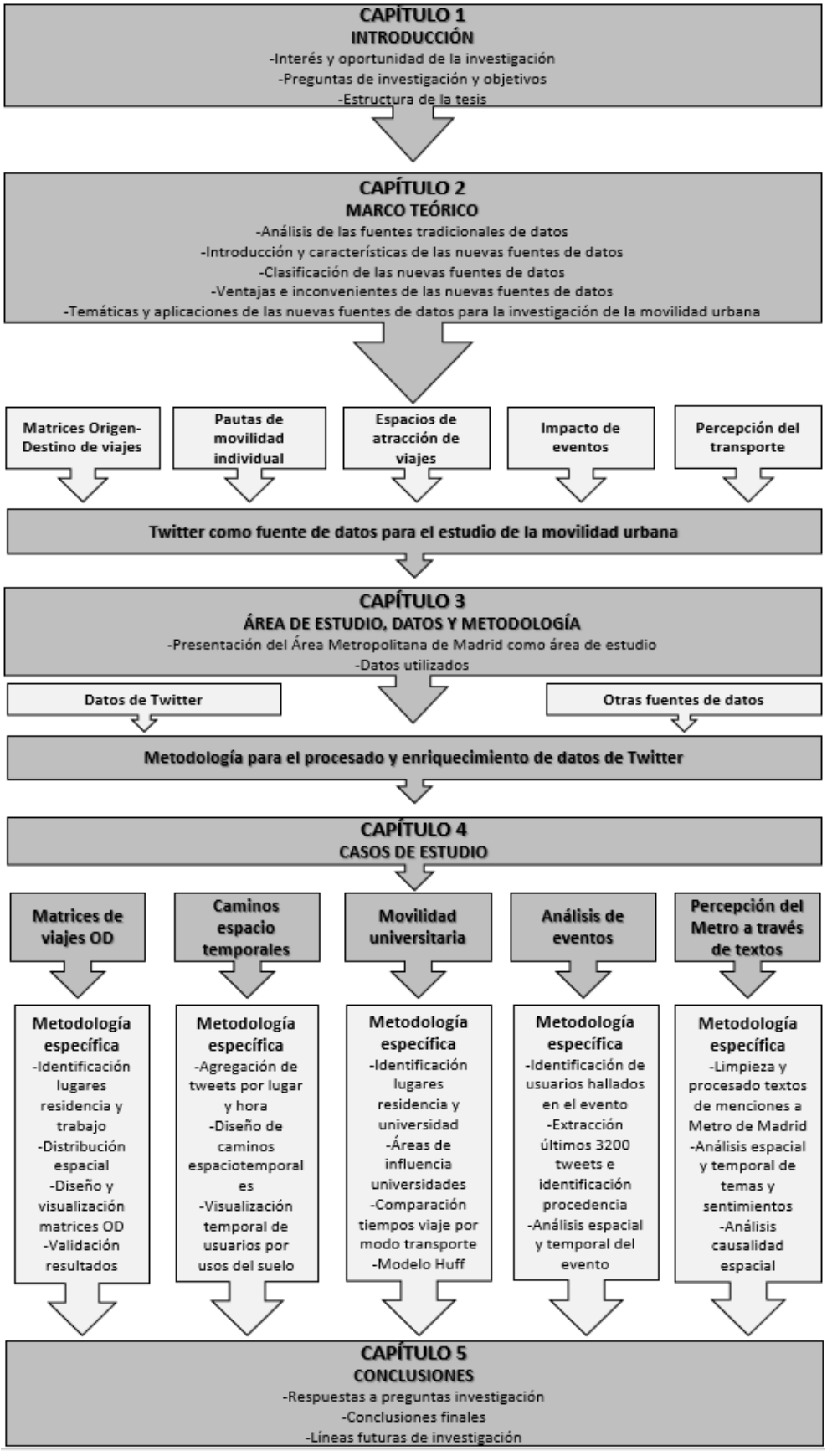
El cuarto caso de estudio investiga la distribución cambiante de la población en la ciudad durante un macroevento como la *World Pride*, y compara los resultados obtenidos frente a los patrones espacio-temporales observados durante una semana habitual. Además, pone en manifiesto una metodología para identificar los lugares de procedencia de los visitantes. Este trabajo se corresponde con el artículo titulado "*Análisis del impacto espacio-temporal de la World Pride 2017 en Madrid a partir de datos de Twitter*", publicado en la revista *Estudios Geográficos* en el año 2020.

El último caso de estudio investiga el valor semántico de los textos de *Twitter* en el campo de la movilidad, a partir del análisis semántico y la extracción de temas y sentimientos sobre un sistema de transporte público muy utilizado como es el caso de la red de Metro de Madrid. Además, busca analizar el efecto de diversas variables como la población o el número de puntos de interés sobre la distribución espacial de los usuarios en la red. Este capítulo se corresponde con el estudio titulado "*Social media semantic perceptions on Madrid Metro system: using Twitter data to link complaints to space*", enviado a la revista *Sustainable Cities and Society* en el año 2020.

Por último, el quinto capítulo discute una serie de conclusiones a partir de los resultados obtenidos a lo largo de la investigación en relación con las preguntas de investigación y los objetivos específicos marcados en el capítulo introductorio. A continuación, se introducen una serie de conclusiones generales que responden al objetivo principal formulado, se marcan las limitaciones encontradas durante la investigación, y se establece una serie de futuras líneas de investigación a seguir en el futuro.

La tesis está redactada casi en su totalidad en castellano. El documento está escrito en inglés en el resumen y en el capítulo final de conclusiones. De este modo, la tesis cumple los requisitos relacionados con el artículo 15 del Real Decreto 99/2011 del documento de la tesis, para la obtención de la mención internacional en el título de Doctor.

Figura 2: Estructura de la tesis doctoral.



Fuente: Elaboración propia.

2. MARCO TEÓRICO

2.1. Fuentes tradicionales de datos

A nivel local, suele producirse un vacío legal sobre las competencias en cuanto a recopilación, tratamiento y difusión de la información sobre movilidad urbana, por lo que las fuentes de datos han dependido generalmente del departamento de tráfico o transporte correspondiente. En las principales ciudades, existen datos de flujos de tráfico urbano. Sin embargo, los datos oficiales disponibles para el investigador suelen ser menores y diferentes entre unas ciudades y otras.

Tradicionalmente se han usado las Encuestas Domiciliarias de Movilidad (EDM) para recopilar datos del origen y destino de los viajes, los medios de transporte empleados y sus etapas, duración, hábitos de transporte de la población y otros datos relevantes (Griffiths, Richardson, & Lee-Gosselin, 2000; Miralles-Guasch, 2012; Miralles-Guasch et al., 2015; Ortúzar S. & Willumsen, 2011). Además, las EDM recogen gran cantidad de datos sobre la población, permitiendo conocer con detalle sus características sociodemográficas y económicas, así como información relevante sobre la vivienda. Estos cuestionarios suelen presentar tres partes: datos sobre el lugar de residencia del encuestado, datos sociodemográficos del individuo, y datos sobre los viajes. Al tratarse de una fuente específica para el estudio de la movilidad y para alimentar el desarrollo de modelos de transporte, la riqueza de los datos obtenidos es amplia.

Sin embargo, las EDM también presentan inconvenientes: su modelización requiere de enormes muestras de hogares encuestados, los datos están adaptados a zonificaciones de transporte, conllevan un considerable trabajo de campo, y un gran esfuerzo personal, lo que eleva los costes para su realización. Estos costes acaban reduciendo la periodicidad de las encuestas y la resolución espacial, ya que en ocasiones se trabaja con zonas de transporte de gran tamaño. Otras limitaciones consisten en que normalmente se obtienen solo datos de viajes en días laborables, y es posible encontrarse con una omisión de viajes, ya que los encuestados tienden a declarar los viajes que consideran más importantes, pero omiten viajes cortos o que no se realizan diariamente, o desplazamientos que consideran menos relevantes. En algunas ocasiones este tipo de encuestas pueden estar algo sesgadas ya que cada vez es más difícil que las personas seleccionadas respondan a los encuestadores, y cuentan con un menor grado de penetración en determinados colectivos sociodemográficos (Gutiérrez-Puebla et al., 2019; Zhao, Ghorpade, Pereira, Zegras, & Ben-Akiva, 2015).

Otro tipo de fuente habitual son las encuestas panel sobre movilidad, que recogen información de una muestra de población durante un periodo determinado de tiempo (por ejemplo, una semana), y en distintas fases (por ejemplo, antes o después de una actuación en la red de transportes). De esta forma se persigue recopilar información sobre cambios en patrones de movilidad, normalmente como consecuencia de algún tipo de actuación en materia de transporte. Nuevamente, la larga duración de este tipo de encuestas eleva su coste de preparación. Por otro lado, también se realizan conteos y aforos en los distintos medios de transporte, como estimaciones estadísticas representativas de su uso. Para ello se obtiene una muestra aleatoria dentro de un ámbito concreto, que es computada estadísticamente, resultando en estimaciones rápidas y de bajo coste. Sin embargo, difícilmente se obtienen datos sobre la población y sus características.

En otros casos se utilizan técnicas cualitativas, como diarios de viaje, observaciones, entrevistas, grupos de discusión, etc. Estas técnicas son métodos más accesibles sobre la movilidad, basándose en la interacción directa con la población objeto de estudio (Pazos, 2005). Se han usado con cierta frecuencia para analizar la movilidad de determinados grupos de población. También se puede recoger información de fuentes indirectas, como los propios censos de población, para obtener datos a partir de cuestionarios sobre los lugares de trabajo y vivienda y del medio de desplazamiento utilizado (García-Palomares, 2010). Aun así, estos censos cuentan cada vez con menor información, su periodicidad de elaboración y publicación es baja y su coste, nuevamente, es elevado.

(Miralles-Guasch & Martínez, 2013) han evaluado la valoración que realizan los profesionales en la planificación y gestión del transporte de las fuentes de datos tradicionales de movilidad urbana. Se muestra cómo, a pesar del interés de estas fuentes, son métodos de obtención de datos poco dinámicos, que requieren de una constante actualización para tener datos adecuados, haciendo que sean fuentes costosas y difíciles de mantener. Además, conllevan un tiempo muy importante de preparación (por ejemplo, con la preparación de las zonificaciones de transportes o las estimaciones de las muestras), que afectan directamente al coste y datos obtenidos. También, la baja resolución tanto temporal y espacial hacen que el tipo y cantidad de datos obtenidos sea a veces limitado, dando en ocasiones como resultado bases de datos incompletas (BITRE, 2014). Por último, las fuentes tradicionales de datos presentan graves problemas de muestreo en espacios específicos como barrios marginales o conflictivos.

2.2. Nuevas fuentes de datos para el estudio de la movilidad urbana

2.2.1. Introducción a las nuevas fuentes de datos

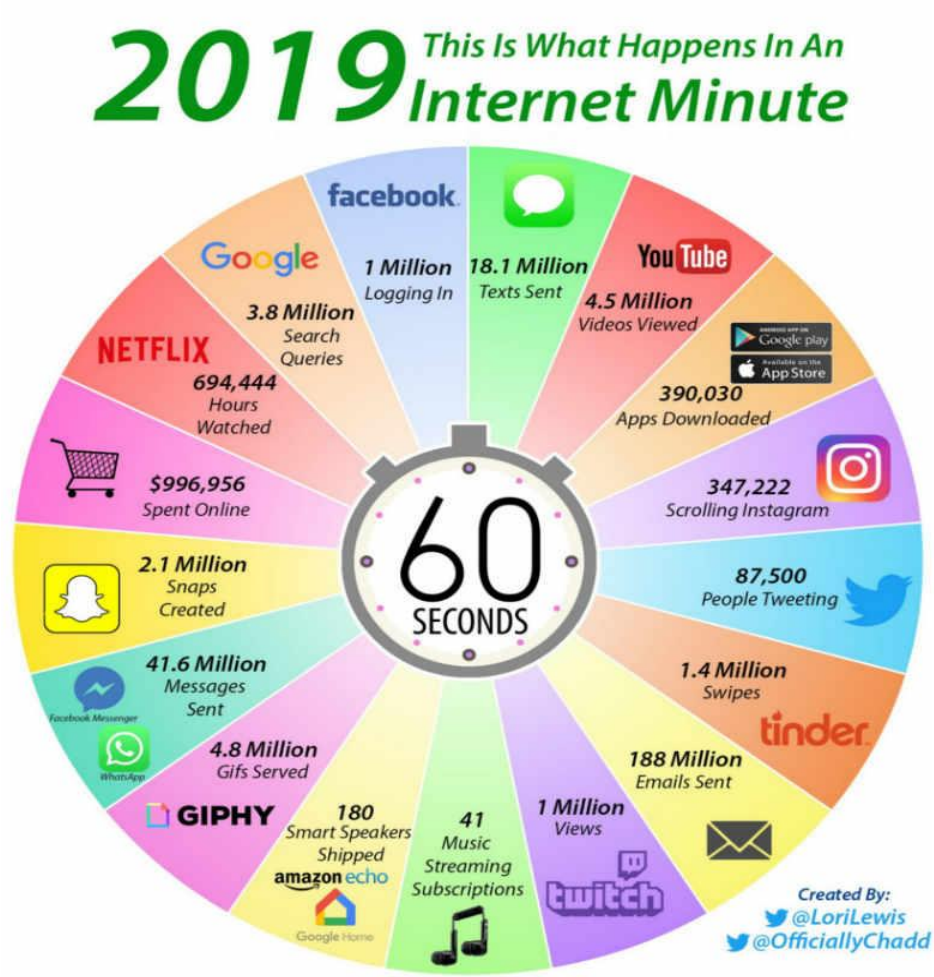
El avance tecnológico ha abierto oportunidades para que las fuentes de datos tradicionales puedan ser complementadas con datos procedentes de otras fuentes, vinculadas al desarrollo de las TIC y las redes sociales, las *Smart Cities*, o el *Big Data*. A la hora de abordar muchos de los temas actuales de investigación, es necesario aunar esfuerzos para acceder a nuevas formas de datos debido a una necesidad creciente de extender la amplitud de fuentes de datos disponibles para la investigación (Lansley et al., 2018). La Información Geográfica Voluntaria presenta un gran interés ya que es creada a partir de los ciudadanos como sensores, recopilada, y diseminada por individuos o grupos en internet de forma gratis y voluntaria (Steiger, de Albuquerque, & Zipf, 2015).

De muchas formas distintas, la actividad diaria de las personas, sobre todo en las ciudades, está monitorizada continuamente por una serie de dispositivos y sensores. El aumento de la disponibilidad de dispositivos móviles equipados con sensores GPS, los ordenadores de alto rendimiento, y las conexiones de internet de banda ancha con servidores avanzados han permitido a los usuarios participar activamente en la creación de contenido cuando interactúan con aplicaciones móviles (Steiger, de Albuquerque, et al., 2015). Cuando una persona realiza una llamada por teléfono móvil, hace búsquedas por internet, envía e-mails, interactúa en una red social, sube o descarga vídeos, usa una tarjeta de transporte o de crédito, o se desplaza usando navegadores GPS, deja un rastro digital de forma voluntaria o involuntaria, una huella formada por datos muchas veces georreferenciados y que se almacenan en diversos repositorios, públicos y privados, por lo que se produce una continua explosión de datos (Gutiérrez-Puebla et al., 2016; Gutiérrez Puebla, 2018). La Figura 3 refleja la cantidad de datos que se genera en un minuto en las diferentes redes sociales y plataformas basadas en las TIC.

Las nuevas fuentes de datos basadas en las TIC son herramientas potentes y dinámicas, fáciles de tratar y que ofrecen la posibilidad de hacer análisis más profundos y complejos, que permiten complementar la información obtenida de encuestas de movilidad. Las TIC engloban un conjunto de herramientas de recogida de datos que pueden ser activas o pasivas dependiendo de si el usuario es consciente o no de la propia generación de

información (Miralles-Guasch et al., 2015). De hecho, la mayor parte de estos datos se produce de forma pasiva y automática (Batty, 2013).

Figura 3: Datos producidos en un minuto por diferentes plataformas basadas en las TIC.



Fuente: (Lewis, 2019).

Se denomina *Big Data* a los datos masivos y de distinta naturaleza producidos a un ritmo vertiginoso por medio de estas nuevas tecnologías. Este término se acuñó por primera vez en las comunidades científicas a finales del siglo pasado, pero no fue hasta el final de la pasada década cuando se volvió popular. Actualmente, el término *Big Data* está muy extendido en internet, pero aún no es un término claramente definido, sino que depende de la perspectiva tecnológica, industrial, investigadora, o académica desde la que se aborda (M. Chen et al., 2014; S. Li et al., 2016). Debido a que muchas de las nuevas fuentes de datos tienen la posibilidad de georreferenciar sus datos, podemos hablar de *Spatial Big Data* (SBD): grandes cantidades de datos que tienen referencia especial y que

pueden ser, por lo tanto, capturados, almacenados, agregados, analizados y comunicados espacialmente.

Se está viviendo una auténtica revolución de los datos, que están adquiriendo un valor cada vez mayor para las empresas y la sociedad. La explosión es tan grande que se están catalogando los datos como el petróleo del siglo XXI. Recientemente los expertos están asegurando que el *Big Data* y su facilidad de ser recogido, almacenado y procesado está configurando una cuarta revolución industrial, la denominada revolución de la información. Hablamos de una revolución que se está fraguando en los últimos años, configurándose como un paso más en la evolución de la ciencia. Con este cuarto paso de la ciencia, la exploración de los datos, los investigadores pasan a interrogar al mundo con instrumentos complejos a larga escala que hacen observaciones en el *Big Data* para proceder y almacenar la información como conocimiento en ordenadores (Hey, Tansley, & Tolle, 2009).

Una geografía basada en los datos está emergiendo en respuesta a la riqueza del flujo de datos georreferenciados. Académicamente el *Big Data* invierte el problema clásico de la muestra donde identificamos una pregunta y recogemos datos para responder a esta pregunta. En su lugar, recogemos primero los datos, y una vez obtenidos, formulamos que preguntas queremos hacer a los datos. Los datos ya no son solo una forma conveniente de calibrar, validar y testear modelos, sino más bien la fuerza motora tras los análisis (Miller & Goodchild, 2014). El *Big Data* puede situarse en el área disciplinar de la teoría y metodología del tratamiento de datos geoespaciales. Además, la geografía posee una poderosa herramienta para el manejo de estos nuevos datos asociados al *Big Data*. Los SIG son una herramienta madura y sin rival para la ciencia de datos por sus habilidades para procesar datos espaciales y no espaciales, aunque no estén perfectamente estructurados, a partir de medios computacionales y visuales. Se puede decir que en el campo de las nuevas fuentes de datos para la geografía, los SIG y el *Big Data* son las "dos partes de un todo" (S. Li et al., 2016). La Tabla 1 recoge el desarrollo de la disciplina geográfica y como se ha llegado a la situación actual, según la cual hoy estaríamos en un cuarto paradigma de la ciencia.

Tabla 1: Evolución de la ciencia desde la geografía.

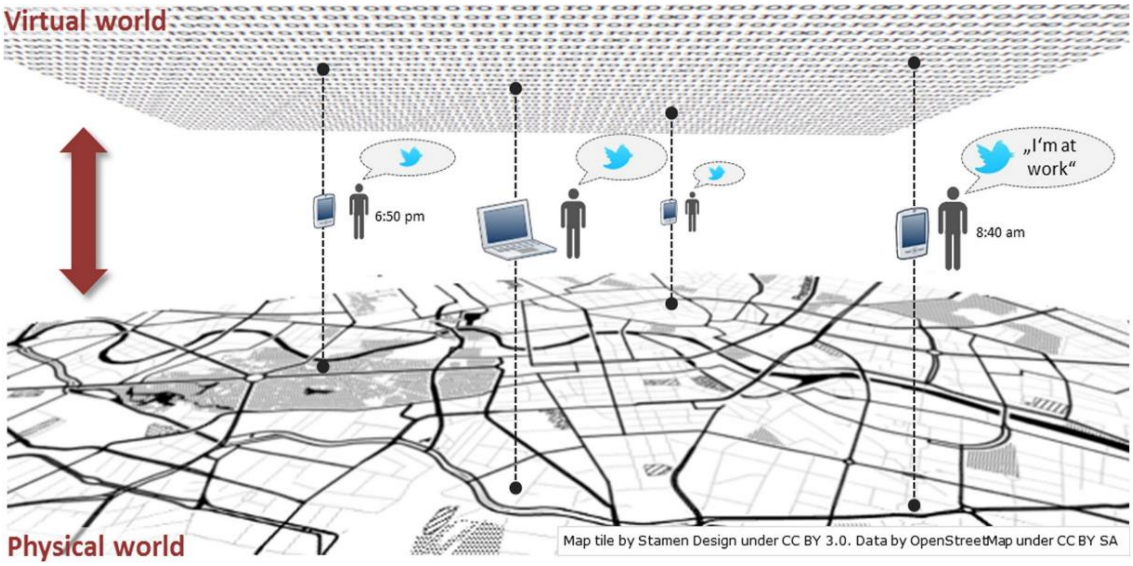
Paradigma de la ciencia	**Tiempo establecido**	**Cuando**	**Definición general**	**Definición desde la geografía**
1º: ciencia empírica o experimental	Milenios	Pre-Renacimiento	Descripción de fenómenos naturales	Primeros estadios de la geografía (desde la geografía descriptiva de la Antigüedad hasta la geografía regional francesa).
2º: ciencia teórica	Siglos	Pre-ordenadores	Formulación de modelos teóricos	Geografía teórica y cuantitativa.
3º: ciencia computacional	Décadas	Pre-*Big Data*	Simulación de sistemas complejos	Desarrollo de los Sistemas de Información Geográfica
4º: ciencia de los datos o exploratoria	Años	Ahora	Exploración estadística de los datos, *data mining*	*Big Data* geolocalizado

Fuentes: (Gutiérrez Puebla, 2018; Kitchin, 2014; Miller & Goodchild, 2014).

El cuarto paradigma de la ciencia consiste en una transición hacia nuevos sistemas basados en la automatización, la robotización, la inteligencia artificial, y el Internet de las Cosas (Bloem et al., 2014; Schwab, 2017). Para entenderlo, hay que ver las TIC como un nodo de enlace entre el mundo virtual, donde se halla la información, y el mundo físico, donde se proyecta dicha información en el espacio y el tiempo (Figura 4). Asociada al surgimiento del *Big Data* surge el concepto de las *Smart Cities*: ciudades sensorizadas por constelaciones de instrumentos, conectadas a través del Internet de las Cosas. Se trata de múltiples redes, que proporcionan datos continuos sobre los movimientos de personas y mercancías y sobre el estado de los sistemas y estructuras que componen una ciudad (Xia, Yang, Wang, & Vinel, 2012). Las *Smart Cities* buscan alcanzar un desarrollo eficiente y sostenible, una participación ciudadana activa, y una mejor calidad de vida. Para ello se basan en el *Big Data*, ya que los datos masivos enriquecen el conocimiento

acerca de las ciudades y ofrecen nuevas oportunidades para la interacción social y la planificación (Batty, 2013).

Figura 4: Las TIC como enlace entre el mundo físico y el virtual.



Fuente: (Resch, Zipf, Beinat, Breuss-Schneeweis, & Boher, 2012; Steiger, Westerholt, Resch, & Zipf, 2015).

2.2.2. Características de las nuevas fuentes de datos

2.2.2.1. Las 3 Vs

El *Big Data* posee tres características principales: volumen enorme de datos, velocidad alta de creación y variedad amplia de fuentes y tipos de datos. Son las conocidas como las *3Vs* (Figura 5) (M. Chen et al., 2014; Kaisler, Armour, Espinosa, & Money, 2013; Kitchin, 2013). En los últimos años numerosas investigaciones están añadiendo otras características, que los lleva hablar de las *4V* o las *5V*. Sin embargo, en este apartado, se entrará en detalle en las tres características que conforman las *3V*, y se explicará al final el carácter optativo de otros dos términos con el que se conforman las *5V*.

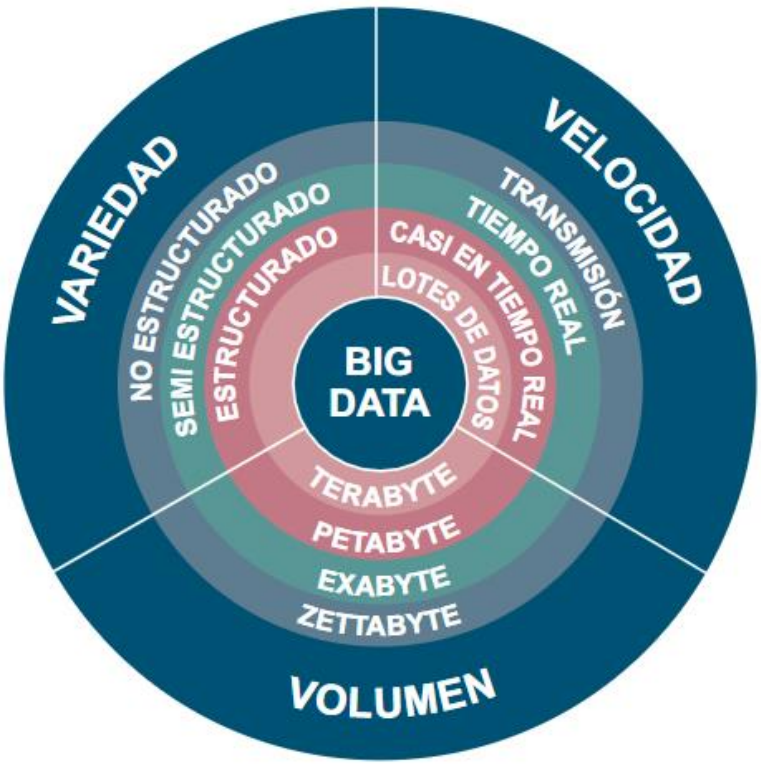
El Volumen hace referencia a la enorme cantidad de datos disponibles. Está directamente vinculado al crecimiento del tráfico de datos producido por la cantidad de dispositivos móviles que utiliza la población mundial, y a la cantidad de datos que poseen las diferentes empresas. Según la compañía tecnológica *Oracle*, el volumen de datos en todo el mundo crece anualmente sobre el 40%. Las unidades de medida son ya terabytes, petabytes o zetabytes. Se recogen miles de millones de registros al día, y estos datos no

se borran, sino que se mantienen ya que sirven de apoyo para analizar procesos y realizar predicciones (Clarke, 2013; Gutiérrez Puebla, 2018).

La Velocidad mide el proceso de la creación, descarga, y agregación de los datos. Está vinculada al ritmo acelerado al que se producen los datos y a la velocidad de las tecnologías y de sus funciones de monitorización y procesado. Los datos se generan de forma constante y continua en tiempo real o casi real, por lo que es posible seguir procesos en streaming y hacer análisis y monitorizaciones en tiempo casi real (Clarke, 2013; Gutiérrez Puebla, 2018).

La Variedad es la medida de la diversidad de tipos, formatos y fuentes de datos. Está asociada al amplio espectro de fuentes desde las que se producen los datos. El *Big Data* se compone tanto de datos estructurados en forma de tablas y bases de datos, datos semiestructurados como ficheros HTML o JSON, y datos no estructurados como archivos de texto, vídeos, audios, imágenes, etc. (Clarke, 2013; Gutiérrez Puebla, 2018).

Figura 5: Las escalas de las *3V* del *Big Data*.



Fuente: (Gutiérrez-Puebla et al., 2019).

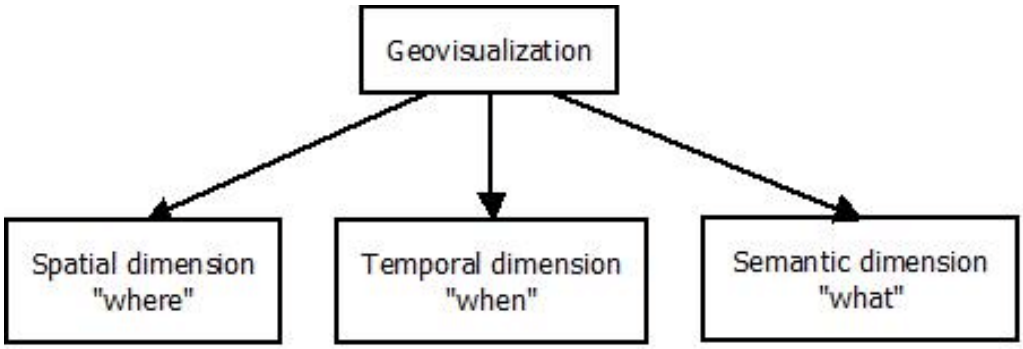
Como se mencionó al principio del apartado, el *Big Data* presenta fundamentalmente tres características o *3V*, aunque sin embargo hay otros dos términos que podrían integrarse y formar las *5V*: la veracidad y el valor. La veracidad está ligada a la fiabilidad, certeza,

calidad e integridad de los datos. El valor es el valor económico que poseen los datos para las empresas. Sin embargo, la veracidad también es una característica intrínseca a los datos tradicionales, es decir, no es un término exclusivo del *Big Data*. Por otra parte, el valor no es una característica con la que cuentan los datos, sino que es un atributo que los propios investigadores confieren a los datos cuando los convierten en conocimiento e información.

2.2.2.2. Las tres dimensiones del Big Data

En los últimos años, el desarrollo de las nuevas fuentes de datos ha permitido una evolución paralela de las herramientas y plataformas de geovisualización en los SIG, usándose frecuentemente estas técnicas para facilitar la identificación e interpretación de patrones y relaciones en datos complejos. Los datos pueden mostrar información espacial, temporal o semántica (Figura 6). Estos tres atributos o dimensiones coinciden también las propiedades básicas de la movilidad humana: el desplazamiento de las personas a diferentes sitios cuenta con características espaciales (localizaciones, distancias), temporales (duraciones, frecuencias en el tiempo), y semánticas o sociales (características demográficas, socioeconómicas, usos del suelo).

Figura 6: Dimensiones de los datos geovisualizados.



Fuente: (Segerberg & Bennett, 2011).

La dimensión espacial se refiere al entorno en el que se generan los datos en un espacio geográfico tridimensional. Esta dimensión responde a la pregunta "dónde", representándose los datos en un espacio geográfico, permitiendo su localización y visualización. La dimensión espacial es la dimensión clave cuando hablamos en término de geolocalización, georreferenciación, o geovisualización (Segerberg & Bennett, 2011), entendiéndose en su concepción cartesiana como plano o contenedor en el que las relaciones sociales ocurren en una geometría de coordenadas xy o de latitud y longitud (Shelton, 2017).

La geolocalización de una gran parte de las nuevas fuentes de datos proviene de dispositivos GPS alojados en teléfonos móviles, que almacenan las coordenadas x e y con un error posicional de unos pocos centímetros. En otros casos, la geolocalización viene dada por un dispositivo fijo que registra los datos, como los sensores o las terminales de chequeo donde se utilizan las tarjetas de transporte público. En términos generales, los datos generados por el primer método disponen de una alta resolución espacial ya que cada dato está localizado por sus coordenadas geográficas, y no por unidades de agregación (Gutiérrez Puebla, 2018).

Al disponer de coordenadas x e y, los datos pueden ser implementados fácilmente en entidades de puntos en un SIG. Con estos puntos se pueden construir por ejemplo mapas de densidades que permiten analizar pautas espaciales con mayor facilidad (Gutiérrez Puebla, 2018). Sin embargo, estos datos cuentan con coordenadas bidimensionales, pero normalmente no tienen una coordenada "z" de altura. Aun así, los datos se pueden representar tridimensionalmente añadiéndoles un valor de altura.

La dimensión temporal está fuertemente conectada con la dimensión espacial, ya que los datos ubicados en el espacio pueden cambiar de localización a lo largo del tiempo. Por tanto, el espacio y el tiempo son muy dependientes el uno del otro, hablándose normalmente del término "espacio-temporal". La dimensión temporal responde a la pregunta "cuándo", pudiendo los datos encontrándose en una secuencia temporal lineal o cíclica. Normalmente, los datos son representados en una secuencia lineal debido a su mayor facilidad para visualizar cambios en el tiempo. Esta secuencia está compuesta por intervalos de tiempo y por puntos o instantes en el tiempo (Segerberg & Bennett, 2011).

Los datos geolocalizados asociados al *Big Data* suelen poseer una considerable resolución temporal que permite la monitorización de procesos espacio-temporales prácticamente a tiempo real. Esto es debido a que cuando se genera un dato, se registra la fecha y hora en la que este dato ha sido creado, teniendo información actual que permite hacer estudios evolutivos y análisis detallados en el caso de movilidad diaria (Gutiérrez Puebla, 2018). Los avances recientes de los SIG permiten, por ejemplo, diseñar visualizaciones y mapas dinámicos o interactivos que muestran los cambios en el tiempo de la distribución de los datos o usuarios en el espacio.

Por último, la dimensión semántica es la representación de los fenómenos en el mundo real para dar significado a los datos. Esta dimensión responde a la pregunta "qué", y se refiere con frecuencia al tema haciendo posible identificar y describir los objetos

espaciales en el territorio. Por tanto, se confiere a los datos de información humana de diversa índole como información socioeconómica, demográfica, política, etc., que da un sentido a lo que se visualiza, ya que lo que se dice de un tema o evento es tan importante como su localización en el espacio-tiempo (Segerberg & Bennett, 2011).

Algunas fuentes de datos como las tarjetas de transporte o bancarias contienen campos de gran utilidad, como información sobre el domicilio del usuario, información demográfica o socioeconómica. Sin embargo, aunque casi toda la totalidad de los datos provenientes de las nuevas fuentes basadas en las TIC cuentan con un valor espacial y temporal, son pocas fuentes las que poseen información semántica, siendo necesario en muchos casos agregar dicho valor semántico a los datos a partir de técnicas de agregación o enriquecimiento.

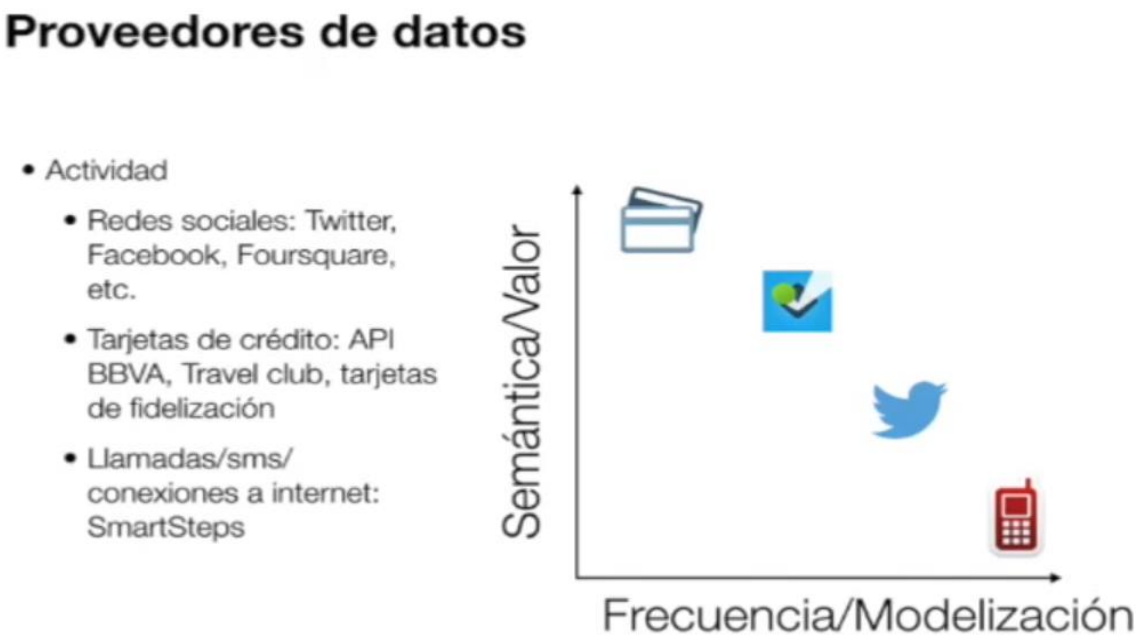
2.2.3. Clasificación de las nuevas fuentes de datos

La constante evolución, mejora y actualización de las TIC han contribuido a generar una diversa gama de nuevas fuentes para el estudio de la movilidad. Las TIC engloban un conjunto de herramientas muy diversas: actividad de telefonía móvil, uso de redes sociales, registros de operaciones con tarjetas de crédito, tarjetas de transporte público, registros de consumo, datos recogidos en tiempo real con GPS, imágenes de cámaras, etc. (Gutiérrez-Puebla et al., 2016). Estas fuentes presentan diferente resolución espacial y temporal entre ellas.

Las fuentes de *Big Data* se pueden clasificar según el modo en el que los datos son generados: por máquinas (sensores activos, GPS, sensores que recopilan datos de forma pasiva, actividad del teléfono móvil captada por antenas de las operadoras de telefonía, etc.), por procesos (datos generados por empresas y gobiernos, como datos de tarjetas bancarias o de transportes), o por fuentes humanas (redes sociales, plataformas como *OpenStreetMap*) (Lansley et al., 2018). En relación con la movilidad, se pueden clasificar los datos respecto a si estudian las infraestructuras, los vehículos o los usuarios. Con las TIC se pueden ver los atributos de las infraestructuras, si están operativas, si necesitan mantenimiento, se pueden formular previsiones, o diseñar mapas dinámicos de su rendimiento. Con los vehículos, es posible monitorizar la localización, el rendimiento, el conductor, o realizar telemetrías. En cuanto a los usuarios, las TIC contemplan su estado, actividad personal, y comportamiento.

En esta tesis se ha empleado la clasificación que (Moro, 2016) realizó de las nuevas fuentes de datos (Figura 7), en función de la frecuencia con la que se producen los datos y el valor de la información que puede obtenerse. Se puede ver que a medida que aumenta la resolución temporal de las fuentes de datos, tiende a decrecer el valor de la información que aportan los atributos de otros campos cualitativos. Así, la geolocalización de llamadas de teléfono o de mensajes de texto en *Twitter* producen datos con una frecuencia muy alta, pero la información temática que puede obtenerse de las mismas es de menor valor. De estas fuentes habitualmente se puede obtener solamente las propias coordenadas, el momento temporal o alguna característica socioeconómica asociada a los usuarios. Por el contrario, las tarjetas de crédito se usan con una frecuencia menor, teniendo por tanto una resolución espacial y temporal baja ya que solo aportan la localización de los individuos en momentos muy puntuales (Gutiérrez Puebla, 2018), pero el valor temático de la información que proporcionan es mucho mayor, con un gran detalle en cuanto a los atributos sociodemográficos y económicos de los usuarios.

Figura 7: Fuentes de información por frecuencia y valor semántico.



Fuente: (Moro, 2016).

2.2.3.1. Teléfonos móviles

La compañía de monitorización *StatCounter* reveló recientemente que por primera vez desde 1980, la plataforma móvil *Android* ha reemplazado al sistema operativo *Windows* como el principal modo de acceso a internet (Simpson, 2017). El mundo virtual está entrando en una nueva era en la que los teléfonos móviles están desplazando a los

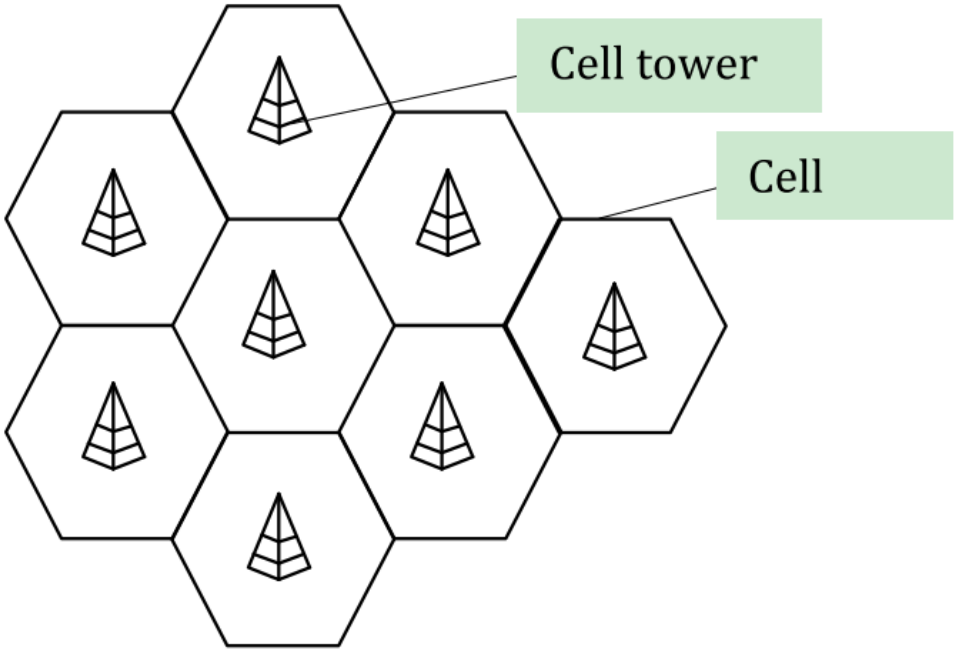
ordenadores como los principales medios de interacción con la sociedad en línea. El acceso a los teléfonos móviles o *smartphones* es cada vez mayor y más sencillo, viviéndose una auténtica proliferación en la que los teléfonos móviles son ya parte indispensable de la vida diaria de la mayor parte de la población metropolitana. Como consecuencia, el aumento del uso de los teléfonos móviles conlleva un espectacular crecimiento de la creación de datos. Un ejemplo de la función de los teléfonos móviles en el comportamiento de la movilidad urbana radica en la incorporación de aplicaciones diseñadas por empresas de transporte para informar al usuario de los tiempos de llegada del vehículo, o de las paradas disponibles en el recorrido a realizar.

Con el aumento de su funcionalidad y complejidad, los *smartphones* cuentan con aplicaciones que sirven para muchas funciones. Entre ellas está la posibilidad de recoger la ubicación del usuario gracias a la función de GPS o la localización de los puntos de la red telefónica a los que se está conectando. En esta segunda opción, la más frecuente, los datos geolocalizados de un teléfono móvil son generados como resultado de la comunicación del dispositivo con una red celular mantenida y operada por operadores en red o antenas (Figura 8). Cuando un usuario realiza una llamada o usa la red 4G de su dispositivo, lo conecta a un operador que determina su posición a partir de la antena y torre a la que se está conectando. Como resultado se crean unos datos llamados *call detail record* (CDR). Estos datos representan la actividad realizada (ya sea una llamada telefónica, un mensaje *sms*, o una sesión de datos), el remitente, la hora a la que se inició la llamada, la duración de la llamada, y las coordenadas x e y de la antena o torre a la que se conectó el teléfono móvil al iniciar la llamada (C. Chen, Ma, Susilo, Liu, & Wang, 2016). Las empresas telefónicas preservan la privacidad de los usuarios, manteniendo siempre su anonimato. Para averiguar las coordenadas x e y del usuario que realizó la llamada se representan las áreas de influencia de las distintas antenas mediante polígonos *Thiessen* (Gutiérrez-Puebla et al., 2016).

La telefonía móvil permite recoger grandes cantidades de datos sobre transporte metropolitano e información sobre la estructura de las áreas metropolitanas y sus propiedades dinámicas (Louail et al., 2014). Estos datos pueden ser utilizados para estudios de movilidad, o para cartografiar la densidad de llamadas a distintas horas del día y analizar la evolución espacio-temporal de la intensidad de las actividades de la ciudad, o visualizar patrones de estructura urbana (Reades, Calabrese, & Ratti, 2009). A partir de la geolocalización de las llamadas, es posible identificar patrones de movilidad

de diferentes grupos de población, las infraestructuras más utilizadas, o los espacios más transitados (Lathia, Smith, Froehlich, & Capra, 2013). El uso de teléfonos móviles para determinar los patrones de movilidad de un individuo conlleva un acercamiento mucho más preciso respecto a fuentes de datos tradicionales gracias a la capacidad de monitorear a tiempo casi real la actividad de una persona (Reddy et al., 2010). En el apartado 2.3.3. de la tesis se entrará en más detalle en la resolución espacial y temporal de los datos de telefonía móvil y se compararán frente a los datos obtenidos a partir de la red social *Twitter*.

Figura 8: Esquema de una red celular de telefonía móvil.



Fuente: (C. Chen et al., 2016).

2.2.3.2. Web 3.0: redes sociales y plataformas web

El concepto de Web 2.0 constituyó un avance importante en el uso de internet, al conferir a la red la interacción y la facilidad de compartir información. Con la Web 2.0 los usuarios ya no eran meros receptores de información, sino que internet se concebía como una plataforma en la que sus usuarios podían generar una enorme cantidad de contenidos (Goodchild, 2007). Desde hace unos años se habla ya de la Web 3.0, cuyos pilares son la interoperabilidad, las tecnologías semánticas y el ambiente de computarización social. La idea básica se halla en definir la estructura de datos y enlazarlos de modo que su uso sea más eficaz y accesible (Aghaei, Nematbakhsh, & Khosravi Farsani, 2012). La obtención

de datos se ha vuelto mucho más sencilla gracias a su disponibilidad en páginas de descarga web proporcionadas por instituciones y empresas, plataformas *OpenData* o Infraestructuras de Datos Espaciales (Gutiérrez-Puebla et al., 2016). La participación ciudadana de forma voluntaria, apoyada en el desarrollo de la Web 3.0, ha contribuido también a la producción de datos masivos geolocalizados. Por ejemplo, *OpenStreetMap* es un proyecto colaborativo consistente en la creación de callejeros digitales de cobertura global a partir de la elaboración y actualización de datos por parte de usuarios voluntarios (Gutiérrez-Puebla et al., 2016). Otro ejemplo destacado es la producción de entradas de información de la plataforma *Wikipedia*.

Una de las fuentes de datos más populares y usadas de internet son las redes sociales, herramientas de comunicación que permiten crear, almacenar, compartir e intercambiar información e ideas en redes y comunidades virtuales con otros usuarios (Cao et al., 2014; Zeng & Gerritsen, 2014). Las redes sociales están conformadas por comunidades de individuos que construyen un perfil online para comunicarse con otros usuarios, compartiendo intereses o actividades comunes (Gutiérrez Puebla, 2018; Steiger, de Albuquerque, et al., 2015). Su función general consiste en la interacción con otras personas a través de diferentes medios, produciéndose de forma diaria una importante cantidad de datos.

En muchos casos la información de las redes sociales está geolocalizada, y tiene un carácter dinámico debido a su constante actualización, lo que permite conocer las pautas de distribución de los usuarios y su movilidad. Además, los datos de las redes sociales cuentan con una muy fácil accesibilidad en comparación con otras fuentes de datos controladas por agencias o corporaciones (Shelton, 2017). Los datos de redes sociales han sido utilizados para estudiar la forma de las aglomeraciones urbanas basándose en la actividad de la gente (Zhen, Cao, Qin, & Wang, 2017), y son útiles para analizar la estructura urbana y la actividad socioeconómica relacionada (Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2017; Yao Shen & Karimi, 2016; P. Zhang, Zhou, & Zhang, 2017). En el caso del estudio de la movilidad, las redes sociales son valiosas para informar acerca del tráfico a distintas horas del día, compartir horarios de las redes de transporte público o eventos en los sistemas de transporte, dar a conocer empresas y modos de transporte, o planificar proyectos que afecten a distintas áreas urbanas.

La red social más usada a nivel global es *Facebook*. Esta plataforma se compone principalmente de un muro o espacio en el que cada usuario puede escribir mensajes, o

subir contenido audiovisual. Este contenido puede ser comentado por otras personas. Además, *Facebook* dispone de otros servicios como mensajería instantánea o páginas donde multitud de personas pueden ingresar para compartir sus gustos sobre un tema en concreto. Sin embargo, los datos de *Facebook* son privados, por lo que el investigador depende de la empresa para poder acceder a ellos. Otra de las redes sociales más populares es *Twitter*, por su sencillez y facilidad para usarse desde cualquier dispositivo. Se hablará con más detalle de esta red social en el apartado 2.3 de la tesis.

Otras redes sociales de gran interés son las basadas en la geolocalización de fotografías, por su facilidad de combinación con otras redes sociales. Las comunidades más importantes son *Instagram*, *Flickr* y *Panoramio*. Las dos últimas permiten visualizar fotografías etiquetadas espacialmente en mapas. Estas fuentes permiten conocer los espacios más fotografiados, los principales puntos de atracción turística y la intensidad de su uso, siendo posible también recoger imágenes de un trayecto y mostrarlas cartográficamente a escala local, habilitando el análisis de la movilidad de los turistas (García-Palomares, Gutiérrez, & Mínguez, 2015; Salas-Olmedo, Moya-Gómez, García-Palomares, & Gutiérrez, 2018). También destaca *Foursquare*, red social que permite compartir con otros usuarios a través del GPS lugares específicos marcados en mapas por el propio individuo, lo que le convierte en un programa ideal para el estudio de puntos de reunión, zonas con mayor tránsito o atracción turística, o actividades basadas en el tipo de equipamiento compartido.

Existen otro tipo de fuentes de datos basadas en la Web 3.0 como las plataformas web para reserva de alojamientos o viajes. Los usuarios de internet emplean páginas como *TripAdvisor* o *Booking* para reservar hoteles, o *Airbnb* para conseguir un alojamiento en la ciudad que van a visitar durante un determinado periodo de tiempo (Gutiérrez, García-Palomares, Romanillos, & Salas-Olmedo, 2017). Los datos generados a partir de estas páginas no solo son valiosos por su información espacial y temporal (los alojamientos de *Airbnb* están geolocalizados en su plataforma web), sino también porque permiten a los usuarios valorar sus viajes o alojamientos, confiriendo a los datos de una muy rica información semántica (Gutiérrez Puebla, 2018).

2.2.3.3. Tarjetas inteligentes de transporte público y tarjetas bancarias

Las tarjetas inteligentes o *smartcards* son un tipo de fuente de datos de gran interés en la investigación de la movilidad metropolitana. Estas tarjetas cuentan con un circuito integrado microprocesador que permiten tanto la identificación del titular como el

almacenamiento de información asociada. Las tarjetas inteligentes de transporte público (Figura 9) son bastante usadas en diversos medios (autobús, Metro o tren), siendo muy valiosas para el estudio de la movilidad gracias a su fácil adquisición, su alta demanda, su constante uso, y su facilidad para almacenar datos sobre viajes de los usuarios. Al ser pasadas por un programa lector, permiten el almacenamiento de patrones de viaje, siendo un modo muy útil de analizar dinámicas urbanas a corto, medio y largo plazo. Entre los datos que se obtienen, destacan los datos del usuario, el medio de transporte utilizado, la parada de origen, y en algunos casos la parada de destino, los trasbordos efectuados, la fecha y el tiempo empleado en el viaje (Cao et al., 2014). Sin embargo, las tarjetas de transporte público no se emplean en todos los viajes (quedan fuera el resto de modos de transporte). Además, en algunos casos solo se registra el lugar y momento en el que se accede al medio de transporte, pero no el lugar y momento de salida, por lo que este ha de ser inferido (Gutiérrez-Puebla et al., 2019).

Figura 9: Tarjeta inteligente de transporte público de la Comunidad de Madrid.

Fuente: Consorcio de Transportes de la Comunidad de Madrid.

Las tarjetas bancarias también permiten ubicar a su portador en los dispositivos y cajeros donde se hayan usado, incluyendo máquinas facturadoras de billetes de transporte público que aceptan pago por dicho medio. Igual que las tarjetas de transporte público, suelen

contar con datos del dueño de la tarjeta, además de datos de cada una de sus compras como la fecha, el lugar o el tipo de establecimiento. La georreferenciación de cada una de las transacciones bancarias realizadas por los usuarios de las tarjetas permite realizar un seguimiento de los individuos y conocer su movilidad en la ciudad (Lenormand et al., 2015). Aun así, la movilidad recogida en las tarjetas bancarias está muy ligada al consumo, lo que causa muchas limitaciones en el análisis de la movilidad urbana (Gutiérrez-Puebla et al., 2019).

2.2.3.4. Fuentes de datos wireless, y navegadores

Otras fuentes de datos basadas en las TIC son los sensores que detectan vehículos o usuarios a través de tecnología *wireless*. Entre ellos destacan sensores acústicos, radares de microondas, sensores de infrarrojos o *LIDAR*, dispositivos *Bluetooth*, o etiquetas de identificación de frecuencia por radio (BITRE, 2014). A su vez, los sistemas de detección de imagen por video se pueden aplicar para realizar conteos de peatones mediante cámaras especializadas, siendo una aplicación muy útil para el estudio de aglomeraciones en determinados lugares o eventos.

Finalmente, otro tipo de dispositivos que produce información de gran interés para el estudio de la movilidad es el compuesto por los sistemas de navegación con los que están equipados hoy en día la mayoría de los propios vehículos, y que se alimentan de datos GPS. Estos sistemas poseen redes viarias muy detalladas con información sobre las características del viario, e indican al conductor en un mapa su ubicación, los servicios más cercanos o el cálculo de rutas óptimas a efectuar en su trayecto, recogiendo a la vez dicha información. Gracias a la base de datos del navegador que almacena las rutas efectuadas es posible hacer una recopilación de trayectos realizados por una persona a lo largo de un tiempo determinado y obtener otros datos de interés como el tiempo empleado. En los últimos años han surgido compañías que utilizan este tipo de fuentes como *Google Transit*, *Uber*, *BlaBlaCar*, u otras muchas similares. Cabe destacar empresas como *TomTom* o *Waze*, que poseen redes que incluyen velocidades de los vehículos en cada uno de los tramos de la red viaria cada cinco minutos, lo que permite realizar análisis dinámicos de accesibilidad considerando la gestión (Moya-Gómez & García-Palomares, 2015).

2.2.4. Ventajas de las nuevas fuentes de datos

Las nuevas fuentes de datos basadas en las TIC pueden recolectar datos de forma constante, frecuentemente en tiempo real, lo que permite la actualización progresiva de la información. Además, la facilidad de disponibilidad y descarga de datos hace que el coste sea normalmente bajo, pudiendo obtenerse y actualizarse información que el propio usuario puede descargar desde internet a través de aplicaciones. Esta información además es en algunos casos de bajo coste o gratuita. Analizar la información disponible en la red es mucho más rápido y tiene menor coste que realizar encuestas (Gutiérrez Puebla, 2018).

Una de las mayores ventajas del *Big Data* es la obtención de bases de datos de alta resolución espacial y temporal. El alto detalle de estos datos permite la monitorización de procesos espacio-temporales prácticamente en tiempo real, la oferta de datos complementarios a las fuentes oficiales y la realización de análisis desde el nivel global hasta de localizaciones concretas (Gutiérrez-Puebla et al., 2016). Desde el punto de vista tecnológico, los datos suelen recogerse de forma pasiva, sin que el usuario tenga que responder un cuestionario, configurar su dispositivo móvil o activar alguna aplicación. Al recogerse de esta manera, los datos muestran lo que la gente hace, y no lo que la gente dice que hace. Además, el usuario no tiene que hacer nada para que los datos se almacenen (Gutiérrez-Puebla et al., 2019).

Con las fuentes de datos tradicionales, los investigadores desarrollaban métodos para recoger muestras representativas y generalizar inferencias de la población que extrajeron en la muestra. Sin embargo, las muestras aleatorias son frágiles y solo funcionan mientras la muestra sea representativa. Otro problema estaba en la dificultad de reutilizar los datos para propósitos más allá de su uso principal (Mayer-Schönberger & Cukier, 2013). Con las TIC, ya no se trabajan con muestras sino con poblaciones: la facilidad para recoger, almacenar y procesar datos digitales significa que, en lugar de trabajar con una pequeña representación de la población, se puede trabajar con la población entera y obtener una cobertura casi total de poblaciones, sistemas, actividades y lugares. Esta ventaja permite por ejemplo realizar estudios comparados entre ciudades (Gutiérrez-Puebla et al., 2019; Miller & Goodchild, 2014). Los usuarios que utilizan las redes sociales pueden ser un buen proxy para estudiar las actividades humanas que ocurren en la superficie de la Tierra (S. Li et al., 2016). Las nuevas fuentes de datos son también capaces de representar la distribución de la población en cualquier momento del día, o de penetrar en barrios marginales de forma efectiva (Gutiérrez-Puebla et al., 2019).

La georreferenciación de los datos de las TIC permite su incorporación en un SIG en entidades de puntos, donde pueden ser tratados y analizados fácilmente con técnicas de geoestadística o estadística clásica (Gutiérrez-Puebla et al., 2016). Estos datos en forma de punto suelen contar en sus metadatos con identificadores con los que se puede agregar, resumir, o representar a individuos u objetos, y obtener medidas de variación e incertidumbre. La amplitud de información y el alto detalle espacial y temporal facilitan el diseño de cartografía y geovisualización de los datos (Ciuccarelli et al., 2014). Estos métodos de análisis y visualización espacial ayudan a los analistas a identificar puntos calientes, clústeres, y valores atípicos en el espacio, y facilitan una comunicación eficaz de los patrones dominantes y relaciones que emergen como resultados del análisis de los datos descargados del *Big Data* (S. Li et al., 2016). Además, los propios SIG se han convertido en una herramienta social aprovechada por las TIC para contar historias y noticias, y compartir mapas y datos geográficos (Sui & Goodchild, 2011).

El *Big Data* tiene además una alta interoperabilidad con las fuentes tradicionales, permitiendo muchas veces crear o incorporar encuestas. Esta complementariedad otorga a las fuentes tradicionales una mayor flexibilidad y un alcance del que no suelen contar. Además, la naturaleza relacional del *Big Data* permite combinar distintas bases de datos. Se pueden añadir nuevos campos fácilmente e incorporar información que las nuevas fuentes de datos no suelen tener como las características socioeconómicas de los viajeros o la identificación del lugar de residencia o trabajo (Gutiérrez-Puebla et al., 2019, 2016).

Las TIC también permiten una amplia flexibilidad en la escala del estudio, ya que, al no haber restricción, los datos no están fijos exclusivamente a una escala residencial. Esto permite a la investigación espacial cambiar del análisis a larga escala de datos agregados al análisis a corta escala de procesos individuales (Gutiérrez-Puebla et al., 2016; Lansley et al., 2018). En contraste con las bases de datos tradicionales, adecuadas principalmente a la extracción de patrones a corto plazo, las nuevas fuentes de datos permiten identificar dinámicas a medio o largo plazo (Long & Shen, 2015).

En el campo de la Geografía, el *Big Data* cuenta con un gran potencial por su capacidad de proporcionar datos abundantes y por la posibilidad de promover los resultados obtenidos mediante técnicas estadísticas y cartográficas a una mayor audiencia (Kitchin, 2013). Las fuentes de datos masivos permiten a los investigadores poner fin a la dependencia de las estadísticas oficiales en diversos campos como la demografía, la actividad económica, la movilidad, los flujos y otros aspectos urbanos (Shelton,

Poorthuis, & Zook, 2015). La extracción de conocimiento de las bases de datos parte de la creencia de que las técnicas estadísticas tradicionales no son capaces de descubrir información oculta en bases de datos masivas, por lo que el reconocimiento de patrones, algoritmos de clusterización, *machine learning*, búsqueda numérica, inteligencia semántica y visualización científica se acomodan mejor al *Big Data* al no requerir los estrictos de la estadística básica (Miller, 2010).

Gracias a las nuevas fuentes de datos y sus análisis en entornos como los SIG, es posible ampliar las miras en estudios de movilidad desde una doble perspectiva. Por un lado, permiten obtener mejores respuestas a problemas ya planteados, con mayor detalle a partir de las posibilidades que ofrece la mayor resolución espacial y temporal de los datos. Por ejemplo, frente a los modelos para estimar pautas de movilidad o de transporte futuros (*forecasting*), han empezado a surgir herramientas apoyadas en *Big Data* para realizar esas predicciones en tiempo real (*nowcasting*) (Hanabusa, 2012). Por otro lado, debido a la posibilidad de adquirir información con atributos sobre diferentes aspectos que no era posible capturar con los métodos tradicionales, es posible ampliar las temáticas de estudio. Surge la oportunidad de estudiar temas que no habían sido tratados o que habían sido dejados de lado ante la falta de información que ofrecían las fuentes tradicionales (Gutiérrez-Puebla et al., 2016). Es el caso de temas como la movilidad turista o la gestión de eventos.

Hay que tener en cuenta que el *Big Data* no va a sustituir en el futuro a las encuestas, sino a complementarlas; las encuestas suministran datos, sobre todo de carácter cualitativo o humano, que el *Big Data* no puede aportar (Gutiérrez Puebla, 2018). Por lo tanto, es necesario contrastar los datos provenientes de las TIC y complementarlos con las fuentes de datos tradicionales. Tanto las fuentes tradicionales como las nuevas fuentes de datos tienen sus propias ventajas e inconvenientes, y su combinación permite realizar un estudio rico en información, tanto cualitativa como cuantitativa y con una alta resolución espacial y temporal (Miralles-Guasch et al., 2015).

2.2.5. Debilidades y retos en el uso de las nuevas fuentes de datos para el estudio de la movilidad urbana

Las nuevas fuentes de datos están cambiando el modo de aproximarnos al estudio de la movilidad urbana, pero su uso en la investigación está aún en sus comienzos; todavía hay

camino por recorrer para su consolidación como fuentes de información de movilidad. Existen barreras que dificultan el uso del *Big Data*, principalmente barreras tecnológicas y de disponibilidad de los datos. Es evidente que estas fuentes presentan inconvenientes, retos que hay que considerar y que provienen de la propia generación masiva de datos. A diferencia de la información procedente de encuestas o censos, la naturaleza desestructurada de los datos condiciona las tareas de almacenamiento y procesado de datos e implica un trabajo previo de depuración, homogeneización o preparación de los datos, por lo que el proceso de salida de los datos es mucho más difícil que la obtención de éstos (Kaisler et al., 2013). Además, los datos basados en las TIC se generan habitualmente con un fin diferente al que se le dan en los estudios de movilidad o del transporte. Se trata por tanto de dar una estructura a los datos y convertir este desorden en información útil (Bosque, 2015).

El principal reto consiste en organizar y gestionar toda la información que compone el *Big Data*, debido a que los datos provienen de fuentes dispersas, diferentes y muchas veces desestructuradas (hay datos con mayor valor que otros), por lo que la información está no normalizada y desordenada. En algunos casos los datos se recogen sin control de calidad, pueden ser irrelevantes o no tener veracidad, careciendo de estructura y metadatos, lo que reduce la precisión de la información (Miller & Goodchild, 2014). Por otro lado, el volumen de datos generados es tan grande que las herramientas y técnicas usadas hasta ahora se muestran inadecuadas a la hora de procesar los datos. El volumen, variedad y velocidad de actualización de las bases de datos originadas a partir del *Big Data* exceden la capacidad de las tecnologías empleadas comúnmente (Shekhar, Gunturi, Evans, & Yang, 2012). Es necesario adquirir tecnología nueva y personal técnico formado en el uso de estas tecnologías, lo que también puede conllevar un coste importante.

Hay dos modos de hacer frente al desorden de los datos: restringir el uso de los datos a operaciones que no intenten generalizar o hacer asunciones sobre la calidad teniendo en cuenta que los datos desordenados pueden ser útiles en las áreas más blandas de la ciencia como la exploración inicial o la generación de hipótesis; o intentar limpiar y verificar los datos eliminando el desorden lo máximo posible para su uso en la construcción del conocimiento científico tradicional (Miller & Goodchild, 2014). Ante esta situación, se establecen tres estrategias para la limpieza, control de calidad y verificación de los datos desordenados: una solución por “multitud”, en la que un gran número de usuarios tiene acceso a los datos y pueden revisarlos; una solución “social”, en la que se implementa

una estructura jerárquica de moderadores voluntarios; y una solución "geográfica o de conocimiento", en la que se preguntan si los datos son falsos o pueden ser falsos según una serie de reglas establecidas (Goodchild & Li, 2012; Miller & Goodchild, 2014). No toda la información almacenada es necesaria para los estudios de movilidad, por lo que es importante seleccionar solamente aquellos campos de información necesarios para el objeto de estudio con el fin de evitar consumir recursos, esfuerzos y tiempo en los procesos de análisis (Gutiérrez-Puebla et al., 2019).

Un problema clave de las nuevas fuentes de datos radica en que suelen presentar ciertos sesgos, ya que las poblaciones con frecuencia son autoseleccionadas en lugar de muestreadas, y muchas veces se desconocen sus características demográficas (Miller & Goodchild, 2014). El acceso a distintos dispositivos móviles, o a redes sociales varía dependiendo de factores como la edad o el poder adquisitivo. El sesgo también varía en función de la fuente: los datos de telefonía móvil presentan un sesgo bajo, mientras que las redes sociales cuentan con un sesgo mayor, debido por una parte a características sociodemográficas (las redes sociales son usadas principalmente por población de 18 a 40 años mientras que a mayor edad hay una menor representatividad), y por otro lado a que una parte importante de los usuarios de las redes sociales hacen uso de las mismas solo de manera esporádica, reduciendo el nivel de representatividad. Cuando se realiza un filtrado para seleccionar los datos más aptos para un estudio, se aumenta el sesgo (Gutiérrez Puebla, 2018). No todos los usuarios son válidos para obtener información de la movilidad urbana por lo que hay que proceder a hacer una muestra de usuarios que tengan datos suficientes y también analizar el ruido que puedan provocar usuarios compulsivos con muchos datos (Gutiérrez-Puebla et al., 2019).

Además, el *Big Data* no registra todas las acciones de la gente, sino solo una parte. No se puede asegurar que los datos obtenidos de las personas sean una descripción precisa de sus vidas completas, sino puede que sean simplemente las vidas que desean presentar en la esfera social (Gutiérrez Puebla, 2018; Miller & Goodchild, 2014). Los viajes a corta distancia, en tiempo breve, en días atípicos o en zonas de baja densidad de población suelen ser subestimados y muy difíciles de ver (Gutiérrez-Puebla et al., 2019). La información sobre las características de los viajes suele ser limitada, por lo que es muy difícil obtener información sobre el modo de transporte, la ruta empleada o el propósito del viaje. Además, las nuevas fuentes de datos suelen tener información sociodemográfica muy limitada. En esta situación, a través del enriquecimiento de datos es posible inferir

información sobre las características del viaje como la hora del viaje, el motivo, modo, ruta o la información sociodemográfica de los usuarios. En este caso es necesario un proceso de validación de los resultados para conocer la claridad del proceso realizado (Gutiérrez-Puebla et al., 2019). También se pueden observar sesgos espaciales: algunas regiones como las localizaciones turísticas o las áreas de recreación se mapean más rápido que otras localizaciones de menor interés (Miller & Goodchild, 2014). Este sesgo también se da a nivel de países, habiendo diferencia entre países con mayor poder socioeconómico, que cuentan con un mayor número de datos e información respecto a países con menores niveles de renta.

La protección de la privacidad, la cuestión de la seguridad, los procesos de distribución de los datos o los derechos de propiedad intelectual son también limitaciones importantes. Cada vez aparecen con mayor frecuencia reglamentos sobre protección de datos que tratan de evitar los riesgos que puedan derivarse del intercambio o mal uso de datos personales. Es muy importante asegurar que tanto en las fases de elaboración como en la presentación o distribución de los resultados no se infrinjan aspectos relacionados con la privacidad de los usuarios (Gutiérrez-Puebla et al., 2019). En muchos casos el acceso a los datos es difícil, al depender directamente de empresas privadas que no comparten los datos o los venden a precios elevados, como por ejemplo, la obtención de datos de tarjetas bancarias (Gutiérrez-Puebla et al., 2016).

El caso concreto del uso de las redes sociales conlleva una serie de barreras. La línea entre la vida personal y la pública de los usuarios a veces es confusa y a la vez conlleva problemas a la hora de tratar la privacidad. La dependencia de gráficos, videos y contenido autogenerado puede producir problemas en cuanto a la accesibilidad, y la propia seguridad de los datos puede estar expuesta por diversas ciberamenazas. En el análisis de la percepción, las empresas pueden mostrar preocupación hacia el criticismo público. Finalmente, conviene considerar que el panorama de las redes sociales puede cambiar en el futuro, pudiendo hacer a veces difícil la adaptación los nuevos cambios. Otros desafios radican en la interoperabilidad con otras fuentes, el acceso para personas con discapacidades o cuestiones multiculturales (Bregman, 2012), el establecimiento de reglas y regulaciones para el control de datos, y el seguimiento de una serie de pasos para diluir la privacidad de los usuarios y aumentar la seguridad de los datos como la anonimización de los usuarios, o la agregación de datos (Kaisler et al., 2013).

Hay que tener en cuenta que el procesado del *Big Data* se ha vuelto dependiente de las tecnologías computacionales (Ash et al., 2018). Algunos desafíos están relacionados con la tecnología necesaria para el tratamiento de datos masivos, como la necesidad de emplear equipos con bastante potencia, de recurrir a la computación en nube, o el uso de bases de datos no relacionales para poder almacenar cantidades enormes de datos. Respecto a los SIG, los principales desafíos son el desarrollo de herramientas para poder tratar, analizar, y cartografiar los datos y la capacidad de los SIG de poder trabajar con bases de datos enormes y de representar capas de millones de puntos.

Académicamente, estamos ante un desafío de calidad vs cantidad en el que se decide que datos son irrelevantes, como asegurar que todos los datos obtenidos sirven para un estudio, cuantos datos son necesarios para hacer un determinado análisis, o la importancia de ganar información acerca de un determinado problema que analizar todos los datos disponibles (Kaisler et al., 2013). Se habla de un retorno al empirismo, en el que el cuarto paradigma de la ciencia analiza grandes volúmenes de datos para obtener respuestas a preguntas científicas sin que sea necesario orientar una investigación por una teoría propia, sino dejar que los datos hablen por si mismos sin formular hipótesis o modelos previos (Kitchin, 2014). Sin embargo, el *Big Data* necesita apoyarse en la teoría, ya que dejar hablar a los datos por sí solos conlleva ignorar conceptos geográficos claves forjados por la ciencia de la información geográfica, y puede llevar a correlaciones sin ningún valor explicativo (Batty, 2013). Además, hay un fallo a la hora de conceptualizar el espacio, debido a que los análisis le dan mucha importancia a las coordenadas de latitud y longitud, pero ignoran el rango de procesos sociales y espaciales atados al dato (Shelton, 2017).

Hay que tener en cuenta que el *Big Data* no está libre de falta de certeza cuando se usa para representar el mundo real: con los datos suficientes, encontrar significados estadísticos es inevitable, pero no necesariamente se puede obtener ninguna información o identificación de relaciones casuales (Lansley et al., 2018). Un desafío científico fundamental es el problema de la generalización, es decir, la eficacia del *Big Data* para proporcionar información en escalas espacio-temporales pequeñas sobre horizontes espaciales y temporales mayores (Miller & Goodchild, 2014). La propia resolución espacial y temporal de los datos es otro aspecto a tener en cuenta junto a la escala: si solo se quieren conocer las pautas de localización de ciertos fenómenos en grandes periodos de tiempo, basta con una resolución alta. Pero cuando se quiere analizar la población en

un sistema urbano en distintos momentos del día, es necesaria una alta resolución espacio-temporal (Gutiérrez Puebla, 2018). La velocidad de los datos es importante ya que los consumidores exigen una visualización rápida, interactiva e interpretable de los datos (Cheng et al., 2016). Por tanto, algunos de los retos a nivel académico consisten en ir más allá de las coordenadas puramente espaciales de los datos, de no centrarse solo en las características espaciales del presente sino examinar procesos espacio-temporales, de no quedarse en lo próximo sino investigar las relaciones de las producciones y flujos de datos, y de no analizar solo los datos generados por las nuevas fuentes de datos, sino contextualizarlos, analizarlos y compararlos respecto a fuentes de datos auxiliares como datos censales (Leszczynski & Crampton, 2016). El trabajo del investigador debe ir más allá del dato geolocalizado para pasar a analizar patrones y procesos espaciales mediante el empleo de análisis geoestadísticos para situar los datos en su contexto socioeconómico.

Como cierre de este apartado, podemos apreciar las ventajas e inconvenientes del uso del *Big Data* cómo fuente alternativa de datos en un ejemplo práctico. El Centro para Análisis Espacial Avanzado de la *University College* de Londres ha estado trabajando con datos de una tarjeta común de transporte público para el área metropolitana de Londres. La información de viajes de los distintos sistemas de transporte público ha sido extraída a partir de los datos de uso de dicha tarjeta. Gracias al uso de las TIC, y al gran detalle de los datos, se ha podido hacer una aproximación que hubiera sido imposible de realizar utilizando las fuentes tradicionales. Sin embargo, la enorme cantidad de datos (1 billón de registros en un periodo de seis meses), el sesgo de la propia fuente (hay un porcentaje de viajeros que se quedan fuera del estudio al no usar la tarjeta de estudio, incluyendo grupos concretos como los turistas), o la dependencia de las propias infraestructuras de transporte (para que la información se registre, hace falta pasar la tarjeta por sensores en paradas; las estaciones con barreras abiertas conllevan un acceso al transporte sin necesidad de usar la tarjeta generándose vacíos de información en determinadas zonas que hay que sortear mediante estimaciones) conllevan una dificultad importante que hay que sortear para poder establecer análisis de movilidad (Batty, 2013).

2.3. *Twitter* como fuente de datos para el estudio de la movilidad urbana

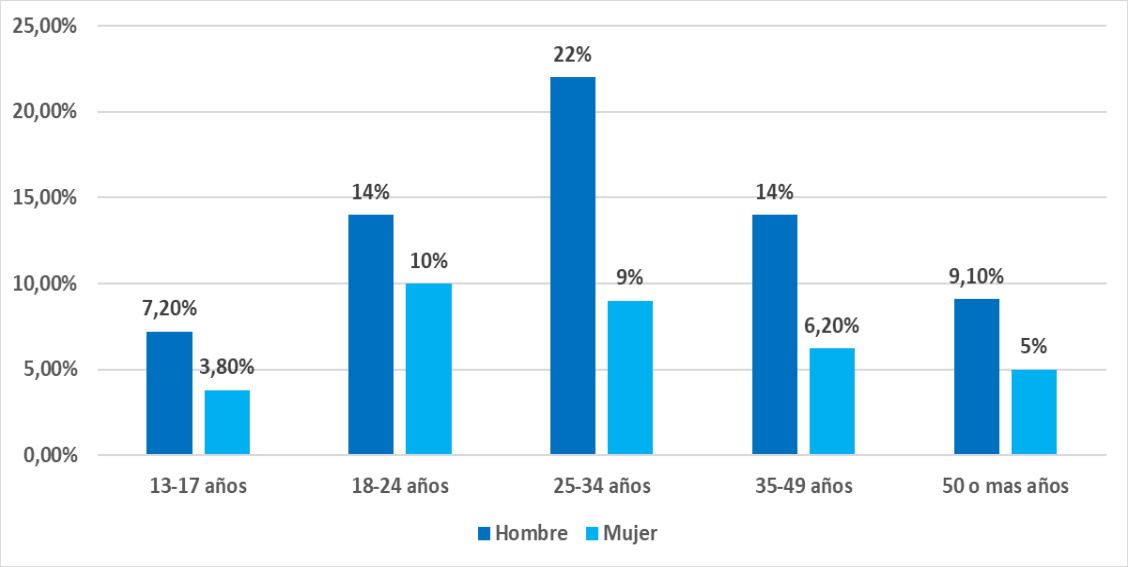
2.3.1. Introducción a la red social Twitter

Previamente hemos comentado que las redes sociales son herramientas electrónicas orientadas a la publicación, relativamente baratas, y abiertamente accesibles, que permiten a cualquier usuario publicar y acceder a la información, colaborar en un esfuerzo común, o construir relaciones en internet. Una de las redes sociales más destacadas es *Twitter*. Esta red social es pública, multimedia, interactiva, de uso gratuito, y escala global. *Twitter* permite a sus usuarios enviar pequeños textos llamados *tweets*, con un máximo de 280 caracteres, además de otros contenidos (Murthy, 2018). Esta aplicación permite establecer comunicaciones digitales de forma rápida, compartir noticias o eventos a tiempo real, opiniones o experiencias, y mantener relaciones temáticas con una sola persona, con un grupo específico de personas o de modo global (Stephens & Poorthuis, 2015; Williams, Terras, & Warwick, 2013). Los últimos años han visto la adopción en masa de *Twitter*, que se ha convertido en parte importante de actividades diarias de millones de usuarios (Longley, Adnan, & Lansley, 2015).

Twitter es una red social inmensamente popular, con aproximadamente 500 millones de usuarios en todo el mundo, y que genera más de 65 millones de *tweets* al día. Se estima que un 80% de estos usuarios publican mensajes desde teléfonos móviles (Ji, Fu, Self, Lu, & Ramakrishnan, 2018). En cuanto a estadísticas demográficas, se calcula que alrededor del 66% de los usuarios son de género masculino, y que un 37% de los usuarios de *Twitter* se hallan en una franja de edad de entre 18 y 29 años, mientras que un 25% de los usuarios tienen entre 30 y 49 años (es decir, más del 60% de los usuarios de *Twitter* tienen entre 18 y 49 años)[3] (Kemp, 2019). La Figura 10 muestra el porcentaje de usuarios de *Twitter* por género y franja de edad. España es el undécimo país del mundo con mayor número de usuarios, estimándose un número de 7.5 millones de usuarios (casi un 20% de la población total mayor de 13 años), de los cuales 140.000 residen en la ciudad de Madrid [4]. Moro (2014)[5] reproduce la movilidad de los usuarios de *Twitter* en España durante un día, realizada a partir de 145 millones de mensajes entre 2012 y 2013, lo que permite apreciar la intensidad del uso de esta red social en el país.

[3] https://www.omnicoreagency.com/*Twitter*-statistics/
[4] https://www.statista.com/statistics/242606/number-of-active-*Twitter*-users-in-selected-countries/
[5] https://vimeo.com/111579945

Figura 10: Porcentaje de usuarios de *Twitter* por franja de edad y género.



Fuente: Elaboración propia a partir de (Kemp, 2019).

Twitter documenta aspectos importantes de las actividades diarias de millones de usuarios a tiempo real, además de sus actitudes sociales y sus opiniones (Longley & Adnan, 2016; Shelton, 2017). Desde su fundación en 2006, el uso de *Twitter* ha ido evolucionando, convirtiéndose en una red de información social, política y económica, y redefiniendo prácticas culturales existentes como escribir un diario o consumir noticias. La facilidad de uso y la naturaleza instantánea de *Twitter* lo han convertido en un medio muy potente para compartir noticias o reportar eventos, desde situaciones mundanas hasta la información emergente sobre política o avisos de emergencias. El objetivo de *Twitter* es que los usuarios respondan a la pregunta "¿Qué está pasando?" (Murthy, 2018). Los usuarios de *Twitter* pueden ser definidos como sensores semánticos con la habilidad de reportar eventos enviando mensajes con huellas digitales (Haghighi, Liu, Wei, Li, & Shao, 2018). Mientras los datos tradicionales pueden tardar horas o días en ser publicados, los datos de *Twitter* son publicados muy rápidamente, permitiendo la captura de un evento en tiempo casi real. Cuando una persona publica un *tweet*, este mensaje es visible de forma instantánea, y cualquier usuario puede responder rápidamente. Compartir estos eventos a una gran comunidad tiene ventajas como aumentar la audiencia del mensaje, movilizar gente a la acción, o permitir a la gente incapaz de atender a un evento a compartirlo en la comunidad (Williams et al., 2013). Desde el sector empresarial, los productores pueden tener acceso ilimitado a la percepción de mercado a partir de comentarios de consumidores en *Twitter* (Ajao, Hong, & Liu, 2015). En definitiva,

Twitter puede ser definida como una "villa global" de individuos que están dando actualizaciones instantáneas en temas y áreas en las que tienen conocimiento o participan a tiempo real: si ocurre algún suceso en algún rincón del mundo, habrá algún usuario que informará de este evento en *Twitter* (Murthy, 2018).

Para entender *Twitter*, hay que entender que son las redes sociales y el *microblogging*. Ya se comentó en el apartado 2.2.5.2 de la tesis en qué consisten las redes sociales. El *microblogging* es una forma de interacción que cumple con los tres conceptos claves del *blogging*: contenidos en forma de post o tema, posts mantenidos juntos por un autor que les confiere de contenido común y controla la publicación, y posts individuales que pueden ser agregados de forma fácil. El *microblogging* se puede pensar como una evolución del *blogging* gracias al aumento de la accesibilidad de una tecnología móvil cada vez más barata, y al desarrollo de la Web 3.0 y las redes sociales. Las características propias del *microblogging* incluyen la facilidad para los usuarios de publicar rápidamente actualizaciones cortas y comentarios de manera regular a tiempo real, dando un método de comunicación innovador que puede ser visto como un híbrido del *blogging*, el mensaje instantáneo y la red social (Williams et al., 2013). Se puede pensar también en el tema o post como un evento. Es en este concepto donde radica una parte del poder seductivo de *Twitter*, ya que los usuarios se sienten contribuidores importantes cuando publican un post o evento sobre un tema o noticia (Murthy, 2018).

Twitter, además de incorporar características del *blogging*, posee elementos propios de los servicios de redes sociales, como la construcción de perfiles de usuario, y la posibilidad de establecer y compartir conexiones con otros usuarios (Williams et al., 2013). Además, las redes sociales como *Twitter* tienen una geografía que mezcla la dimensión digital con lo material. Los usuarios de *Twitter* establecen vínculos sociales basados en intereses comunes en vez de un lugar compartido. Sin embargo, *Twitter* retiene una conectividad local fuerte, haciendo posible comprender la influencia de la distancia espacial en la conectividad entre usuarios (Stephens & Poorthuis, 2015).

Los *tweets* son mensajes totalmente públicos y accesibles. Estos mensajes pueden ser enviados por internet, desde cualquier ordenador o dispositivo móvil. Cualquier teléfono móvil, incluso el modelo más básico, es compatible con *Twitter*, por lo que la tecnología es potencialmente accesible en todo el mundo. La facilidad de enviar rápidamente un *tweet* desde cualquier teléfono móvil, la baja curva de aprendizaje para usar *Twitter* (es una red social relativamente fácil de usar por cualquier persona), y el tiempo requerido

para publicar un *tweet* (tiempo mínimo en comparación con la publicación de cualquier otro material en internet) son tres conceptos que han contribuido al enorme crecimiento de esta plataforma en los últimos años (Murthy, 2018).

Una ventaja importante de la publicación de *tweets* desde teléfonos móviles consiste en que los *tweets* pueden ser geolocalizados a partir de coordenadas GPS de longitud y latitud del dispositivo utilizado. Por tanto, cuando un *tweet* es producido, *Twitter* graba la información geográfica del usuario en ese momento temporal, junto a una variedad de metadatos, generando datos de grano fino que permiten la monitorización de procesos espacio-temporales a tiempo casi real (Frias-Martinez, Soto, Hohwald, & Frias-Martinez, 2012; Gutiérrez Puebla, 2018). Estos datos se pueden trabajar fácilmente con un SIG al poder incorporarse en entidades de puntos. Procesando los datos por el identificador del usuario, se puede tener una aproximación de la huella digital de dicho usuario al poder observar los diferentes lugares en los que ha estado el usuario a lo largo del día (Gutiérrez-Puebla et al., 2016). Un *tweet* georreferenciado representa una observación del mundo real y cuenta con información tanto espacial como temporal y semántica (Steiger, Lauer, & Ellersiek, 2014). Por tanto, los datos de *Twitter* se han convertido en una de las primeras fuentes globales de datos libres y de fácil acceso, con millones de registros digitales de la actividad humana en el espacio y en el tiempo (Hawelka et al., 2014; Kocich & Horák, 2016).

Los datos de *Twitter* están altamente desagregados y diferenciados tanto en el espacio como en el tiempo, permitiendo una gran libertad en los análisis geográficos. El estampado temporal es muy preciso y permite a las localizaciones temporales estar relacionadas con puntos ubicados en el espacio. Este alto detalle espacio-temporal hace posible la monitorización de patrones de actividad diarios de las personas e investigar dinámicas de concentración, dispersión, y segregación de grupos, haciendo posible desarrollar la geografía urbana más allá de las fuentes tradicionales de datos (que recogen datos de forma poco frecuente y pertenecen principalmente a las características nocturnas de áreas residenciales, dándose una carencia de datos sobre las actividades de los ciudadanos durante el día) (Longley & Adnan, 2016; Longley et al., 2015). Otras fuentes sociales como *Flickr* no son capaces de realizar estos niveles de análisis a causa de tener una granularidad espacio-temporal mucho más baja (Gutiérrez Puebla, 2018).

Twitter es la red social más utilizada en estudios urbanos, no solo por tratarse de una plataforma de fácil alcance y por la posibilidad de descargar los datos a tiempo casi real,

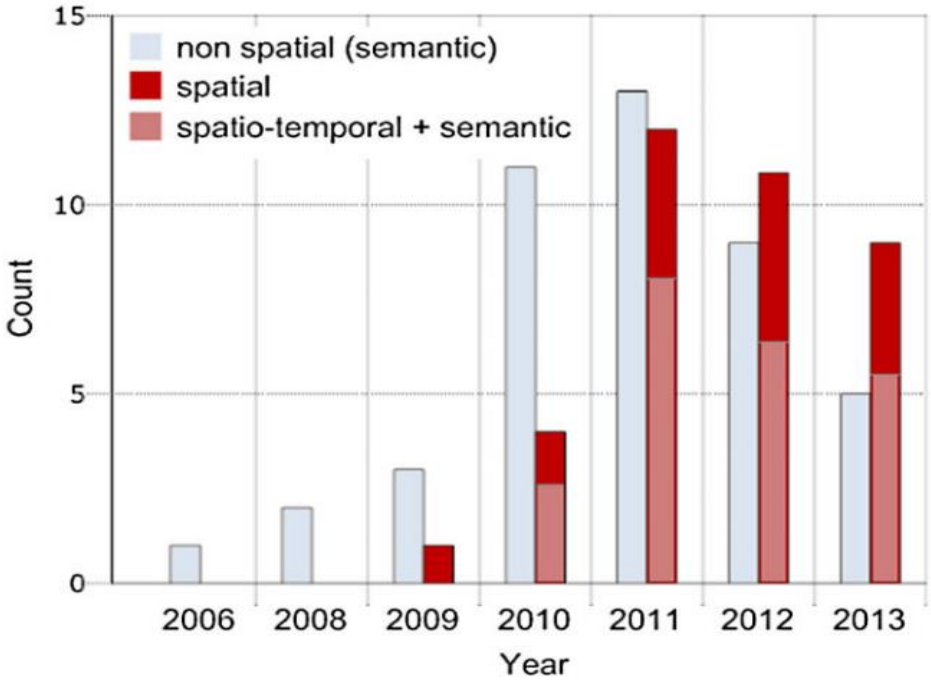
sino también porque estos datos están disponibles de forma gratuita, mientras que los datos de otras redes sociales como *Facebook* o *Instagram* no son accesibles (Gutiérrez-Puebla et al., 2016; Moya-Gómez, Salas-Olmedo, García-Palomares, & Gutiérrez, 2017; Murthy, 2018). El potencial de *Twitter* ha sido reconocido por cada vez un mayor número de campos de investigación en los últimos años. Una explicación al aumento de la investigación de *Twitter* por parte de la geociencia y las ciencias sociales se halla en el aumento del radio de penetración y del uso de *Twitter* por un mayor número de gente que comparte su localización gracias al aumento de la disponibilidad de teléfonos móviles con GPS (Steiger, de Albuquerque, et al., 2015). La investigación académica estudia los datos de *Twitter* a partir de patrones desarrollados a diferentes escalas, principalmente a escala urbana.

Twitter se ha vuelto una herramienta ideal y original para estudiar la movilidad y los patrones de comportamiento de los flujos de un área metropolitana. Gracias a la localización del *tweet*, es posible conocer los cambios de la distribución de la población en la ciudad a lo largo de un tiempo determinado, o las características de la estructura urbana a partir del perfil horario, pudiendo establecer comparaciones entre diferentes lugares o fechas (Gutiérrez-Puebla et al., 2016). A la vez, el acceso instantáneo a opiniones, críticas y datos a tiempo real se antoja de un valor clave a diferentes sectores que usan *Twitter* como herramienta para conducir encuestas y estudiar la opinión de usuarios o grupos sobre el funcionamiento de los modos de transporte (Luong & Houston, 2015). Otros campos de investigación a partir de datos de *Twitter* son el uso de espacios a partir del análisis de movimientos sociales o concentraciones, o el seguimiento de fenómenos en tiempo real como macroeventos o catástrofes naturales (Gutiérrez-Puebla et al., 2016). De hecho, los textos de *Twitter* están siendo explorados como indicadores para sistemas de localización a tiempo real y prevención de enfermedades y de desastres como terremotos, o para ayudar a promover respuestas de emergencias a crímenes (Ajao et al., 2015; Steiger, de Albuquerque, et al., 2015).

(Steiger, de Albuquerque, et al., 2015) han analizado el tipo de investigaciones que han empleado datos de *Twitter*. Según estos autores casi el 50% de los trabajos publicados que están basados en datos de *Twitter* se hallan en el ámbito de la ciencia computacional. Casi un 33% del total están en el campo de la ciencia de la información. En menor grado hallamos investigaciones ubicadas en el ámbito de la geociencia o ciencias de la tierra (7%), o en el campo de las ciencias sociales (menos de un 5%). En cuanto al tema, casi

el 50% de las investigaciones están dedicadas a la detección de eventos (la mitad de ellos sobre eventos de desastres naturales, y solo un tercio sobre control de tráfico). Más del 25% de los trabajos no tienen un contexto específico, mientras que menos de un 15% tienen como foco la inferencia de ubicación a partir de datos geolocalizados. En relación con los metadatos usados, aproximadamente un 33% de las investigaciones aúnan la información espacio-temporal y semántica. Casi el 60% de las investigaciones tratan solo la información semántica, mientras que un 10% emplea solo datos espacio-temporales. Sin embargo, los estudios que han usado datos espacio-temporales han ido aumentando y superando en número a los estudios basados en los datos semánticos en los últimos años (Figura 11). Finalmente, el área de estudio de la mitad de los trabajos de investigación se ubica en Estados Unidos. Otros países a destacar como áreas de estudio son Reino Unido y Japón.

Figura 11: Número de publicaciones por año según la información de *Twitter* empleada.



Fuente: (Steiger, de Albuquerque, et al., 2015).

2.3.2. Estructura y características de los datos de Twitter

Las redes sociales basadas en datos geolocalizados como *Twitter* permiten enlazar el mundo físico y la dimensión digital a partir de tres capas de información: una capa social (o capa de información del usuario), una capa geográfica (la localización del dato a partir

de coordenadas u otros metadatos), y una capa de metadatos semánticos (el contenido del dato) (Steiger, de Albuquerque, et al., 2015). Cuando se realiza una investigación a partir de datos de *Twitter*, hay cuatro aspectos a considerar: el mensaje o *tweet* (el texto con los metadatos asociados), el usuario (los metadatos asociados al perfil del usuario), la tecnología (las APIs utilizadas) y el concepto de la investigación a realizar (el motivo para el que se quieren usar los datos) (Williams et al., 2013).

Twitter se puede considerar como un gran objeto digital que contiene una recopilación de objetos digitales: los *tweets*. Para poder obtener los *tweets*, se utiliza una API (*Application Programming Interface*). Las API son un conjunto de instrucciones creadas por desarrolladores para interactuar con algún tipo de tecnología. *Twitter* ofrece una API abierta que permite a desarrolladores externos crear sus propios programas basados en la API principal, y desarrollar soluciones basadas en los datos de *Twitter* para promover el uso del producto (Kocich & Horák, 2016). Un usuario de *Twitter* puede tener acceso a la API y poder descargar sus datos a partir de una serie de códigos llaves internos que se autogeneran. Los *tweets* son recopilados a tiempo real en la API oficial de *Twitter*, mediante un proceso de *streaming* o descarga continua. Esta API permite la posibilidad de descargar determinados *tweets* a partir de filtros como nombres de cuentas o palabras clave, e incluso permite descargar solamente *tweets* geolocalizados (Steiger, de Albuquerque, et al., 2015).

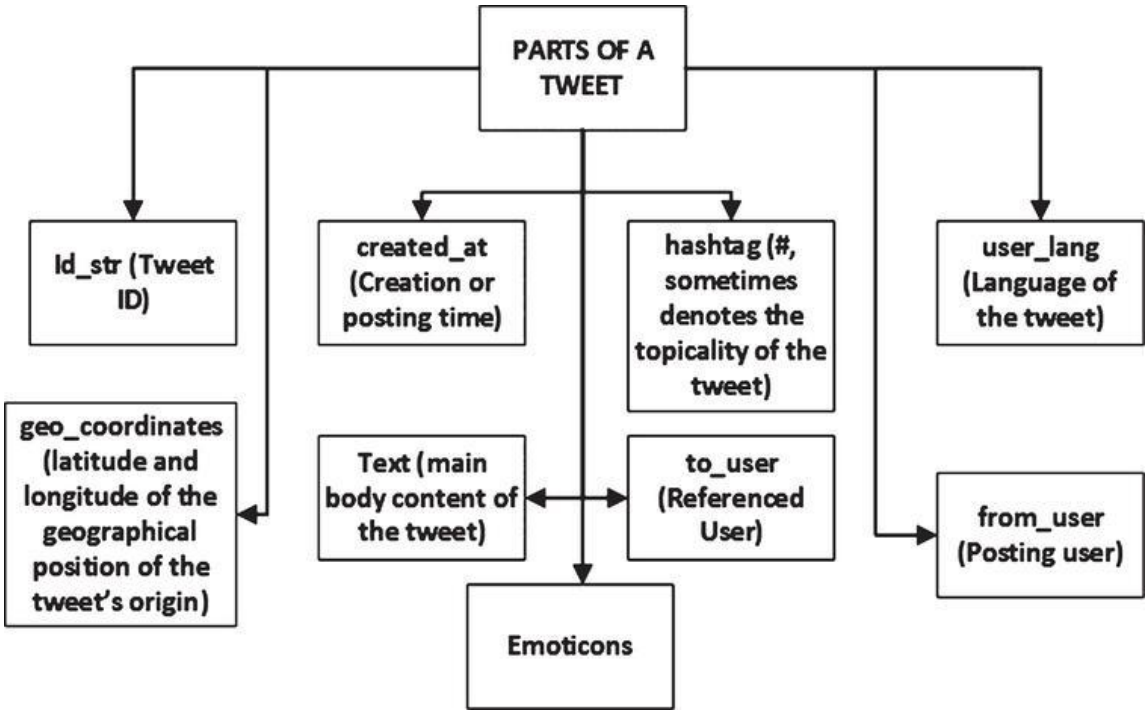
Los *tweets* se descargan en formato JSON (*JavaScript Object Notation*) como objetos con una serie de metadatos. Cada *tweet* cuenta con un autor, un mensaje, una ID única, una estampa temporal sobre cuando fue publicado, y en ocasiones, metadatos de coordenadas geolocalizadas compartidas por el usuario. Además, los *tweets* cuentan con metadatos del usuario que ha publicado el mensaje (como el nombre del usuario de *Twitter*, una ID única de usuario, un número de seguidores, o una biografía de la cuenta). Con cada *tweet* se generan también objetos de entidad: contenidos externos como *hashtags*, menciones, hipervínculos, o contenido multimedia, también con metadatos propios (*Twitter*, 2019). La Tabla 2 y la Figura 12 muestran los principales metadatos que componen un *tweet*.

Tabla 2: Metadatos de un *tweet*.

Atributo	**Tipo**	**Descripción**	**Ejemplo**
Id	Text	Identificador del *tweet*	1099026866227372033
User_id	Text	Identificador del usuario	974022020
lang	Text	Identificador del idioma del *tweet*	es
text	Text	Texto del *tweet*	#retencion nivel amarillo en #a7 (pk 215 al 211 creciente) fuengirola #dgtmalaga #dgt https://t.co/h2tcbyvtcq
Created_at_localtime	Date	Fecha en la que se publicó el *tweet*	2019-02-22T20:23:18
POINT_X	Double	Coordenada de longitud del *tweet*	-4,6082983
POINT_Y	Double	Coordenada de latitud del *tweet*	36,58005905
Shape	Geometry	Geometría del *tweet* en el SIG	Point Z

Fuente: (*Twitter*, 2019).

Figura 12: Estructura de un *tweet*.



Fuente: (Murthy, 2018).

Se puede pensar en el texto como el núcleo del *tweet*. Como se ha mencionado previamente, los *tweets* son mensajes que permiten un máximo de 280 caracteres (hasta el año 2018 el límite estaba en 140 caracteres). Esta limitación requiere de una brevedad en la escritura, dando alza a un diccionario informal de palabras solamente usadas en *Twitter*. Además, cuando un usuario escribe un mensaje en *Twitter*, tiende a incluir muchas abreviaciones (tanto estándar como no estándar), errores tipográficos, emoticonos, y *hashtags*. (Ajao et al., 2015). Por tanto, los *tweets* son mensajes desestructurados, pero, aunque presentan un límite de caracteres, cuentan con una gran versatilidad al poder admitir una amplia gama de elementos, como emoticonos, hipervínculos enlazados, encuestas, imágenes, y contenido multimedia (imágenes animadas, o vídeos). Todos los *tweets* de un usuario pueden ser agregados a partir del perfil de usuario o *timeline*, ya que cada usuario cuenta con un número identificador propio de su perfil (Murthy, 2018).

Twitter cuenta con un método simple pero eficaz para conectar los mensajes con temas mayores, gente específica y grupos. Los *hashtags* son una parte integral de la habilidad de *Twitter* para enlazar las conversaciones a partir de diferentes *tweets* que presentan un tema en común. Este enlace consiste en una o varias palabras claves precedidas por un símbolo #. Este método permite la identificación de temas populares o *trending topics*.

Además, *Twitter* permite que los usuarios puedan reenviar *tweets* a sus seguidores, mediante el *retweet*, un método que facilita la redistribución de *tweets* y *hashtags* fuera de la red más cercana de una persona a audiencias más grandes y desconocidas (Murthy, 2018).

Un *tweet* representa una señal espacio-temporal (unas coordenadas y una estampa temporal presentes como metadatos del *tweet*) en una capa de información semántica (el texto del *tweet*) (Steiger, de Albuquerque, et al., 2015). Todos los *tweets* cuentan con una estampa temporal que indica de forma precisa la fecha y hora en la que fueron publicados. Además, como se ha comentado antes, cada *tweet* contiene datos de coordenadas de latitud y longitud adquiridas a partir de un sensor GPS en el móvil. Aunque no todos los *tweets* cuentan con estos datos espaciales ya que dependen de si un usuario permite la geolocalización de sus mensajes. Pese a que existen algunos métodos para discernir referencias geográficas de *tweets* no georreferenciados, enfocarse solo en los *tweets* geolocalizados asegura un cierto nivel de certeza ya que los *tweets* han sido creados en el lugar desde el cual son geolocalizados (Shelton, 2017). Otros metadatos de utilidad son los atributos del idioma del *tweet* o de la ubicación (no georreferenciada) del usuario. Estos metadatos pueden ser útiles para entender las geografías de los flujos y para revelar patrones e información económica, social, y ambiental (Graham, Hale, & Gaffney, 2014).

Un primer paso en la exploración de datos de *Twitter* a partir de sus metadatos o atributos es usar las coordenadas y la estampa temporal de los *tweets* de un mismo usuario (o grupo de usuarios) para trazar la huella digital del usuario. Un segundo paso consiste en la computación de minería de textos que puedan permitir la caracterización de lugares y la distribución de palabras claves en grupos locales a partir de análisis semánticos y extracción de temas o sentimientos (Pallares-Barbera & Masala, 2016).

2.3.3. Datos de Twitter vs datos de telefonía móvil

Para comparar el granulado de los datos de *Twitter* frente a los datos de telefonía móvil, vamos a analizar las características espaciales y temporales de ambas fuentes de datos en conjunto. Empezando por la dimensión temporal, los registros de los teléfonos móviles tienen una resolución temporal muy alta, con un registro producido en un intervalo de pocos segundos. El número de registros diarios producidos por los usuarios de telefonía móvil es mucho mayor que en el caso de las redes sociales, cuyo uso es más esporádico

a lo largo del día: la frecuencia temporal de *Twitter* está ligada a la propia plataforma, es decir, cada vez que un usuario publica un *tweet*, se graba la fecha y hora en la que se generó el mensaje. También hay que comentar que las redes sociales presentan una gran variabilidad en cuanto a uso, por lo que redes sociales frecuentemente usadas como *Twitter* o *Facebook* cuentan con una alta resolución temporal frente a redes sociales más esporádicas como las aplicaciones basadas en fotografías. Comparando el periodo temporal necesario para obtener muestras de datos adecuadas, con los datos de telefonía se puede trabajar normalmente con datos de 2 o 3 meses. Sin embargo, con las redes sociales, hace falta muestras de periodos mayores (por ejemplo, hace falta periodos de 1-2 años para obtener muestras útiles de *tweets* geolocalizados). Por este motivo, los registros CDR de telefonía móvil son la fuente más usada para estudios temporales debido a su alta penetración y resolución temporal (Gutiérrez Puebla, 2018).

Sin embargo, si hablamos de la dimensión espacial (la dimensión principal a la hora de trabajar con datos georreferenciados), la situación cambia. Los datos georreferenciados de *Twitter* presentan muy alta resolución espacial debido a que cada *tweet* que se publica tiene vinculadas unas coordenadas x e y a partir del dispositivo GPS del teléfono móvil, por lo que se tiene la ubicación del usuario donde publicó el mensaje, con un error estimado de unos pocos centímetros. Al poseer coordenadas de latitud y longitud, los datos de *Twitter* se pueden transferir directamente a un SIG, y generar mapas de puntos con los que analizar pautas espaciales con facilidad o hacer procedimientos puntos-en-polígono para obtener información de usos del suelo (Gutiérrez Puebla, 2018; Longley & Adnan, 2016).

Por el contrario, los datos provenientes de los CDR de telefonía móvil tienen menor detalle espacial al ser la geolocalización menos precisa. Tal como se vio en el apartado 2.2.3.1. de la tesis, el posicionamiento del dispositivo móvil a partir de las antenas de telefonía móvil hace que las coordenadas x e y de los CDR no sean del usuario que utiliza el teléfono móvil, sino de la antena desde donde se conecta el móvil, por lo que es necesario generar áreas de influencia de las distintas antenas de telefonía móvil mediante polígonos de *Thiessen* o *Voronoi*. En algunos casos, aunque no es lo frecuente, se puede estimar la ubicación del usuario mediante técnicas de triangulación (con un error de unos 50 o 100 metros). Como resultado, muchas veces no hay un ajuste adecuado con la distribución de los usos del suelo o las zonificaciones de transportes necesarias para una buena modelización (Järv, Tenkanen, & Toivonen, 2017). Además, la densidad de las

antenas de telefonía es mucho mayor en las ciudades que en las áreas rurales, por lo que el error a la hora de estimar la exactitud posicional puede ser bastante mayor en lugares con una pequeña cantidad de antenas (Gutiérrez Puebla, 2018). Por tanto, *Twitter* cuenta con una ventaja crucial respecto a los datos de telefonía móvil, ya que solo los datos con coordenadas geográficas precisas son válidos para análisis de grano fino como es la escala urbana. Estos datos basados en puntos dan ventajas sustanciales precisamente por no estar limitados por unidades convencionales de área como es el caso de los datos basados en los CDR de telefonía móvil. Los puntos individuales pueden ponerse en relación con una variedad de métodos como la agregación a unidades de área mayores para encontrar concentraciones de *tweets*, o filtrando a partir de otros metadatos disponibles en cada *tweet* individual (Shelton, 2017).

No cabe olvidar la dimensión semántica de los datos. Aquí los datos de telefonía móvil tienen la ventaja de poder disponer de información sobre el usuario, ya que las operadoras tienen mediante contratos datos como la edad, género o domicilio de los clientes, aunque esta información se usa en pocos casos por las limitaciones de privacidad (Gutiérrez-Puebla et al., 2019). Los datos de *Twitter* no cuentan con información acerca de las características del usuario por lo que estas características deben ser inferidas u obtenidas a partir de otras fuentes secundarias, como datos del censo o del catastro (Longley & Adnan, 2016). En cambio, respecto al valor semántico, los datos de telefonía móvil suelen contar con un valor bastante escaso al tener muy pocos metadatos. Mientras, los datos de *Twitter* presentan un valor semántico mayor al disponer tanto del texto del *tweet* del que se pueden extraer datos, como de otros metadatos de interés como el idioma del *tweet*. Con estos datos, es posible indagar en algunos aspectos sociodemográficos del usuario, o extraer sus opiniones y percepciones sobre un determinado tema. Sin embargo, aunque los datos de *Twitter* poseen atributos semánticos, hay que resaltar que son bastante escasos en comparación con datos de tarjetas bancarias o de transporte (aunque en contraposición, la resolución espacial y temporal de este tipo de fuentes de datos es pobre en comparación con la resolución de los datos de *Twitter*).

Por último, teniendo en cuenta la disponibilidad y accesibilidad de los datos, tal como se ha mencionado con anterioridad, los datos geolocalizados de *Twitter* son una fuente global de datos libres y de fácil acceso y descarga. En cambio, la alta fragmentación del mercado de telecomunicaciones móviles impide la disponibilidad de datos de telefonía móvil a escala global. Además, por lo general, estos datos no son libres ni tienen acceso

fácil al depender de diferentes compañías telefónicas (Hawelka et al., 2014). Sin embargo, hay que tener en cuenta que los datos de telefonía móvil cuentan con muestras de usuarios muy grandes, mientras que las muestras de usuarios de *Twitter* son relativamente menores (teniendo en cuenta solamente *tweets* geolocalizados, pueden ser significativamente menores debido al bajo número de usuarios que activan la función de geolocalización en sus perfiles), y más sesgadas, ya que los usuarios de estas redes sociales están concentrados en determinados grupos sociodemográficos.

2.3.4. Debilidades de Twitter como fuentes de datos

Pese a las ventajas que hemos visto que tiene *Twitter* como fuente de datos para los estudios de movilidad, hay que tener en cuenta que también tiene una serie de debilidades y desafíos propios a tener en cuenta (además de otros desafíos relativos al *Big Data* ya comentados en el apartado 2.2.5. de la tesis como la cuestión de la privacidad de los datos). La primera de ellas es que, aunque dispongamos de muestras de datos gratuitas de forma accesible, en realidad la API gratuita de *Twitter* solo puede descargar una muestra del 1% de los *tweets* en el periodo temporal prescrito. Además, los datos georeferenciados son una porción de la cantidad total de mensajes útiles disponibles (Pallares-Barbera & Masala, 2016). La mayoría de los usuarios de *Twitter* no activan la opción de geolocalización de sus mensajes, por lo que podemos decir que una muestra de datos geolocalizados consiste en el 1% de la muestra del 1% de datos que se pueden descargar (Graham et al., 2014; Longley & Adnan, 2016). Las muestras que usan *tweets* geolocalizados para tratar un tema determinado pueden ser de tamaño muy pequeño, y esos *tweets* suelen estar mezclados con una cantidad mayor de *tweets* no relevantes para la investigación. Una solución alternativa para conseguir muestras mayores de datos consiste en descargar *tweets* no geolocalizados y extraer información del campo de localización del usuario que publicó el mensaje, o del propio texto de los *tweets*. Sin embargo, hay que tener en cuenta que más del 85% de los usuarios de *Twitter* dejan en blanco el campo de localización de su perfil, o incluso aproximadamente un 35% de los usuarios ponen una localización falsa (Ajao et al., 2015). Además, tampoco hay uniformidad en la escala (usuarios que publican una ciudad como localización, una región, o un país). En cuanto a la extracción de localizaciones del texto, depende del tipo de investigación que se esté llevando a cabo, y del propio formato de texto (utilización de abreviaciones, por ejemplo).

Una de las críticas más empleadas sobre el uso de *Twitter* es la falta de representatividad o sesgo de la red social (Shelton, 2017). Mientras que los datos de telefonía móvil presentan un nivel de sesgo bajo, las redes sociales cuentan con un nivel de penetración menor y son más utilizadas por grupos específicos, tendiendo a tener un nivel de sesgo mayor. Aunque se tengan datos digitales muy ricos de usuarios, estos usuarios distan de representar al conjunto de la población, dándose un desajuste entre la muestra y la población. En el caso de *Twitter*, las poblaciones mayores de 54 años y las que no cuentan con estudios universitarios están infrarrepresentadas, mientras que la población joven está sobrerrepresentada (Gutiérrez Puebla, 2018; Jurdak et al., 2015). Además, no hay medios confiables de extrapolar los perfiles de los individuos a la población, para así obtener perfiles demográficos. Al estar restringido el rango de los atributos de los individuos en *Twitter*, no se puede obtener un perfil demográfico de los usuarios (Longley et al., 2015).

Otro tipo de sesgo que hallamos en los datos de *Twitter* es de índole geográfica, ya que se desconoce si los usuarios de *Twitter* envían mensajes desde localizaciones específicas. Los usuarios de *Twitter* exhiben una mayor preferencia a retornar su localización más popular que los usuarios de telefonía móvil, por lo que se pierde información de otras posibles localizaciones y se desconoce el impacto en los patrones de movilidad (Jurdak et al., 2015). La solución más óptima es estimar las localizaciones de los usuarios a partir de las coordenadas de localización que generan los dispositivos móviles, pero como se ha comentado previamente, solamente el 1% de los usuarios de *Twitter* envían *tweets* geolocalizados (Ajao et al., 2015). Además, los *tweets* están sesgados de forma desproporcional en áreas urbanas (especialmente en los centros urbanos y de ocio), dándose pocos datos en zonas periféricas o rurales (Shelton, 2017).

A la hora de trabajar con datos de *Twitter*, hay que tener en cuenta que una parte importante de sus usuarios hacen uso de esta plataforma de forma esporádica, reduciendo el nivel de representatividad de la red social. Como consecuencia, la resolución temporal está muy influida por valores extremos, como usuarios con muy baja actividad de uso en sus dispositivos móviles, por lo que una práctica habitual es seleccionar a los usuarios con una mayor resolución temporal. En este sentido, un riesgo a tener en cuenta es la propia sobreestimación de usuarios compulsivos, que dejan muchas más huellas digitales por día que los usuarios medios. Para evitar esta sobreestimación, se puede hacer una agregación espacial y temporal de los datos de forma que tengamos el número de usuarios en cada lugar y franja horaria y no la cantidad de actividad (*tweets*) generados por esos

usuarios. Esto es fácil de hacer gracias al número identificador de *Twitter* que posee cada usuario (Gutiérrez Puebla, 2018).

En cuanto al valor de los datos, el principal desafío radica en la limitación del contenido semántico causada por el propio vocabulario y jerga específicas de *Twitter* y la cantidad limitada de caracteres, que influyen en que los textos sean no convencionales y desestructurados (Ajao et al., 2015; Jurdak et al., 2015). La gramática propia de los *tweets*, junto con el uso frecuente de sarcasmos y palabras irónicas, dificulta el análisis semántico de los textos, por lo que son normalmente tratados como ruido ya que las herramientas de proceso de idioma naturales no tratan bien estos textos (Ajao et al., 2015). El contenido de los *tweets* permanece descontextualizado a no ser que se encuentren modos de darle un valor geográfico, como por ejemplo métodos de inferencia de localizaciones a partir de los textos (Graham et al., 2014). La mayoría de estudios que usan datos de *Twitter* procesan la información textual de los *tweets* mediante filtros basados en palabras como posible solución a estos desafíos (Steiger, de Albuquerque, et al., 2015).

Respecto al análisis y representación de los datos, los principales desafíos radican en que los mapas que tratan datos de *Twitter* suelen quedarse en las coordenadas, sin buscar la riqueza del contenido que puede observarse en esos datos para responder preguntas sobre el fenómeno que se cartografía en cuestión. Uno de los principales problemas es el *overplotting*: mapeado de muchos puntos que se ponen simplemente uno encima de otro en un mapa, haciendo imposible discernir patrones espaciales útiles. Una solución a este problema consiste en la aplicación de una variedad de métodos de recolección de datos, y análisis estadísticos y cartográficos para filtrar ruidos asociados a la densidad de puntos y dar una mayor comprensión de los procesos socio-espaciales que están vinculados con la conectividad de la gente y los lugares (Shelton, 2017).

El creciente cuerpo de investigaciones que usan *Twitter* no es claramente visible y no es fácil de localizar. Por ejemplo, es difícil identificar a primera vista métodos aplicados para el análisis espacio-temporal de datos de *Twitter*. Aunque se están desarrollando metodologías para el tratamiento de los datos, todavía se puede observar una falta de métodos comunes para adaptarse a los datos de *Twitter*. Además, los métodos espaciales actuales solo incorporan los efectos de la escala geográfica de modo marginal. Por último, las limitaciones relacionadas con la incertidumbre de los datos hacen que sea difícil validar y comparar los resultados obtenidos con otras fuentes de datos de referencia (Steiger, de Albuquerque, et al., 2015).

2.4. Temáticas y aplicaciones de las nuevas fuentes de datos para la investigación de la movilidad metropolitana

En la última década hemos vivido un prolongado aumento del uso del *Big Data* para crear bases de datos geolocalizadas con las que estudiar diversas problemáticas. Las aplicaciones en la investigación son muy variadas, cubriendo de forma amplia diferentes campos como geomarketing, movilidad, turismo, diferenciación social, etc. (Gutiérrez-Puebla et al., 2016). Paralelamente, hay trabajos teóricos centrados en el reto y las oportunidades del uso del *Big Data* para la investigación geográfica (Ash et al., 2018; Barnes, 2013; Batty, 2013; Bosque, 2015; C. Chen et al., 2016; Cheng et al., 2016; Gandomi & Haider, 2015; Graham & Shelton, 2013; Gutiérrez-Puebla et al., 2016, 2019; Gutiérrez Puebla, 2018; Kitchin, 2013, 2014; S. Li et al., 2016; Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2019; Miller, 2010; Miller & Goodchild, 2014; Shelton, 2017; Steiger, de Albuquerque, et al., 2015). A continuación, este apartado cita cinco líneas principales de trabajo, que están vinculadas a los casos de estudio desarrollados en la tesis.

2.4.1. Aforos y matrices de viajes Origen-Destino

La gran frecuencia de datos que proporcionan las TIC es de gran utilidad para obtener aforos y mediciones de matrices de viajes, y para poder diseñar matrices de viajes Origen-Destino (OD). Los primeros trabajos de este tipo se basaron en datos de telefonía móvil, y se usaron para estimar velocidades de desplazamiento y tiempos de viaje (Bar-Gera, 2007), para localizar puntos de anclaje residencia-trabajo (Ahas, Silm, Järv, Saluveer, & Tiru, 2010), o para calcular el tráfico de la red viaria en carreteras que no disponen de estaciones de aforo convencionales (Noelia Caceres, Romero, Benitez, & Del Castillo, 2012), mostrando como resultados que las mediciones realizadas tienen buenos ajustes en relación al aforo real. Además, gracias a la resolución temporal que tiene la actividad de los teléfonos móviles es posible identificar los periodos en los que los individuos permanecen en un lugar y los viajes entre lugares que realizan, haciendo posible elaborar modelos predictivos (De Domenico, Lima, & Musolesi, 2013) o generar matrices OD que cuantifican el volumen de viajes (Alexander, Jiang, Murga, & González, 2015; Bonnel, Hombourger, Olteanu-Raimond, & Smoreda, 2015; N. Caceres, Wideberg, & Benitez, 2007; García-Albertos, Picornell, Salas-Olmedo, & Gutiérrez, 2019; Kung, Greco, Sobolevsky, & Ratti, 2014; Louail et al., 2015; Toole et al., 2015). Estos trabajos han

usado datos de telefonía móvil para sus investigaciones aprovechando el importante tamaño de las muestras y la alta frecuencia de datos (Picornell et al., 2015). Sin embargo, el uso de los datos de telefonía tiene algunos inconvenientes, que pueden cubrirse en parte utilizando otras fuentes como los datos de redes sociales (Wu, Zhi, Sui, & Liu, 2014).

Aunque la mayoría de los trabajos dedicados al conteo de aforos o la construcción de matrices OD han usado como fuente los datos de telefonía móvil, en los últimos años han surgido también investigaciones basadas en datos de redes sociales, principalmente *Twitter*. (Perez, Dominguez, Rubiales, & Lotito, 2015) desarrollaron un método para actualizar matrices basadas en tráfico de vehículos en el área metropolitana de Buenos Aires a partir de un almacén de *tweets* previamente filtrados y clasificados por distintos periodos temporales. En Los Ángeles, (Gao et al., 2014) investigaron la eficacia de *Twitter* como herramienta para crear matrices OD comparando los resultados con los datos de la *American Community Survey* mediante un coeficiente de correlación de Pearson. Para ello, se enfocaron en detectar trayectorias individuales y agregaciones de viajes a lugares (usando zonas de tráfico como zonas de origen y destino). También en Los Ángeles, y a partir de este trabajo de Gao et al., (2014), (J. H. Lee, Goao, & Goulias, 2015) buscaron validar la veracidad de *Twitter* cómo fuente, utilizando una aproximación donde extrajeron relaciones OD en días se semana (en un periodo de dos días), agregados por zona espacial, y las validaron frente a otras fuentes (usando un modelo de auto regresión de *Tobit* con datos del censo y de modelos de demanda de transporte tradicionales). La fiabilidad de este tipo de datos ha sido validada en el trabajo de (Lenormand et al., 2014), quienes compararon datos de *Twitter* con información de redes telefónicas y datos oficiales (censos) en las ciudades de Madrid y Barcelona, concluyendo que las tres fuentes de información ofrecen resultados comparables.

Los trabajos pioneros que usaron datos de telefonía para el diseño de matrices OD obtuvieron el número de viajes entre zonas en ciertos corredores de transporte (N. Caceres et al., 2007). En una segunda fase, los trabajos estuvieron enfocados en calcular matrices de viaje con movilidad residencia-trabajo (Ahas et al., 2010; Alexander et al., 2015). Algunos autores fueron más lejos, obteniendo matrices de viajes que incluyen otros tipos de viajes. Este es el caso de (Picornell et al., 2015) quienes trabajaron con categorías de viaje como “otros viajes frecuentes” y “viajes infrecuentes”. Para obtener matrices de viaje, estos trabajos tenían que determinar, como mínimo, las residencias y lugares de trabajo de los individuos. Para este fin, utilizaron la localización más frecuente durante

horas de trabajo, y la localización más frecuente durante horas de tarde-noche. Una vez que las matrices eran obtenidas, algunos trabajos condujeron procesos para expandir los datos de la población en su conjunto, y para verificar los resultados. Métodos similares a aquellos usados en telefonía fueron utilizados en trabajos previos que utilizaron datos de *Twitter* (Lenormand et al., 2014).

La tabla 3 muestra trabajos previos que han utilizado matrices de viajes con datos de telefonía móvil y *Twitter*. Los datos de telefonía son la fuente de *Big Data* utilizada con mayor frecuencia en estudios de movilidad, y *Twitter* es la principal alternativa. Por último, otras fuentes de datos basadas en TIC han también empleadas para la medición de aforos y el diseño de matrices OD han sido las videocámaras (Finnis & Walton, 2007), los sistemas de navegación de vehículos (Dewulf et al., 2015; Hanabusa, 2012; Martin, Jordan, & Roderick, 2008), o las tarjetas inteligentes de transporte (Tao, Rohde, & Corcoran, 2014).

2.4.2. Pautas y recorridos de movilidad individual

La alta resolución espacio-temporal de las nuevas fuentes de datos permite dar un salto del análisis general de la movilidad metropolitana al estudio de la movilidad individual. Los datos de *Twitter* han sido utilizados para el análisis de las pautas de distribución de la población durante el curso del día (Ciuccarelli et al., 2014). (Longley & Adnan, 2016) realizaron un trabajo similar en Londres trabajando con población identificada por etnias y grupos sociales. Recientemente, (García-Palomares et al., 2018) analizaron las distribuciones horarias de los usuarios de *Twitter*, combinando los datos de la red social con datos de usos del suelo para obtener la actividad principal de los usuarios en diferentes momentos del día. Esta información permite además obtener una aproximación a las zonas de origen y destino de los viajes. (Salas-Olmedo & Rojas Quezada, 2017) han cartografiado los patrones de movilidad individual en los espacios públicos de la ciudad de Concepción, en Chile, y a partir de los flujos obtenidos encontraron áreas potenciales de exclusión social.

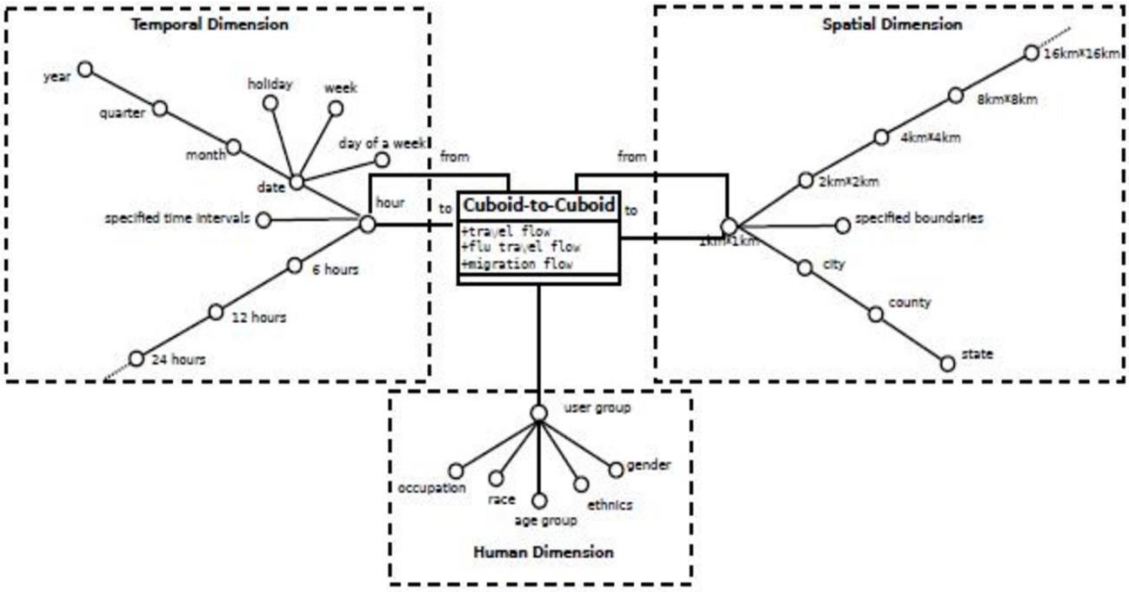
Tabla 3: Aproximaciones a la movilidad a partir de datos de telefonía móvil y de *Twitter*.

Fuente de datos	**Autores (año)**	**Detección de hogar/trabajo**	**Datos de usos del suelo**	**Expansión de los datos**	**Verificación de los datos**
Datos de telefonía móvil	(N. Caceres et al., 2007)	No	No	Si	No
	(Ahas et al., 2010)	Si	No	No	Si
	(Louail et al., 2015)	Si	No	No	No
	(Alexander et al., 2015)	Si	No	Si	Si
	(Bonnel et al., 2015)	No (ya suministrados)	No	Si	Si
	(Toole et al., 2015)	Si	No	Si	Si
	(Picornell et al., 2015)	Si	No	Si	Si
	(García-Albertos et al., 2019)	No	No	Si	No
Datos de *Twitter*	Gao et al. (2014)	No	No	No	Si
	(J. H. Lee et al., 2015)	No	Si (Business Establishments)	No	Si
	(Perez et al., 2015)	Si	No	Si	No
	(Salas-Olmedo & Rojas Quezada, 2017)	No	No	Si	No
	(J. Yin, Soliman, Yin, & Wang, 2017)	No	No	No	No
Datos de telefonía móvil y *Twitter*	(Lenormand et al., 2014)	Si	No	Si	Si

Fuente: Elaboración propia.

Para estudios con diversos grupos sociales, el concepto de información tridimensional es una herramienta eficaz, definiendose la dimensión temporal cómo la duración y distribución horaria de las actividades individuales, la dimensión espacial cómo las configuraciones localizables de los patrones de actividad humana, y la dimensión humana cómo las interacciones sociales con otros individuos. Partiendo de este teorema se han desarrollado herramientas para el análisis espacio-temporal de la localización social como el *datacube* (Figura 13). Esta herramienta consiste en una base de datos tridimensional donde se pueden examinar las dinámicas de un individuo en las redes sociales a múltiples niveles de escalas espacio-temporales y a diferentes reglas de agregación. Una vez construido el *datacube* se puede proceder a la construcción de trayectorias individuales de los usuarios de *Twitter* y a la visualización de patrones espacio-temporales en un SIG, permitiendo también crear un interfaz de cartografía en internet para representar las dinámicas de movimiento a partir de las trayectorias creadas (Cao et al., 2014).

Figura 13: Esquema de un *datacube*.



Fuente: (Cao et al., 2014).

Otra herramienta para el estudio de la movilidad individual es el camino espacio-temporal. Aunque han habido trabajos previos que han mostrado caminos espacio-temporales a partir de análisis de redes de transporte o encuestas (B. Y. Chen et al., 2013; J. Chen et al., 2011; Demšar & Virrantaus, 2010; Fang, Shaw, Tu, Li, & Li, 2012; Farber, Neutens, Miller, & Li, 2013; J. Lee & Miller, 2018; J. Y. Lee & Kwan, 2011; Miller, Raubal, & Jaegal, 2016; Ren & Kwan, 2007; Shaw, Yu, & Bombom, 2008; Tong, Zhou, & Miller, 2015; Yu & Shaw, 2008), hay pocos trabajos en los que los caminos han sido

diseñados a partir de nuevas fuentes de datos. Entre estos trabajos, podemos destacar a (Yue Shen, Kwan, & Chai, 2013), que utilizaron datos de navegadores GPS para ilustrar patrones de desplazamientos en la ciudad de Beijing; (Kang et al., 2010) que a partir de datos de telefonía móvil buscaron representar patrones de movilidad individual en una gran ciudad de China; (Keskin, Çelik, Doğru, & Pakdil, 2014) que diseñaron una aplicación móvil para obtener datos espacio-temporales de 10 participantes de la Universidad Técnica de Estambul con los que visualizar sus caminos dentro del campus; o (Farber, O'Kelly, Miller, & Neutens, 2015) que combinaron datos de telefonía móvil con datos GPS en el área metropolitana de Detroit (EEUU) para medir el potencial de interacción social a escala metropolitana y desarrollar una metodología para comprender los impactos.

En cuanto al uso de datos de *Twitter* para análisis temporales, hay varios trabajos realizados (Antoine, Jatowt, Wakamiya, Kawai, & Akiyama, 2015; Birkin, Harland, Malleson, Cross, & Clarke, 2014; Blanford et al., 2015; Ciuccarelli et al., 2014; García-Palomares et al., 2018; Netto, Pinheiro, Meirelles, & Leite, 2015), pero solo hay constancia de un trabajo que haya empleado esta red social para la visualización de caminos espacio-temporales: (Q. Huang & Wong, 2015), quienes visualizaron caminos espacio-temporales a distintas escalas temporales en distintas ciudades de Estados Unidos con el fin de mostrar la posibilidad de crear caminos espacio-temporales a partir de datos de *Twitter*.

Otros trabajos han empleado otras fuentes de datos distintas, pero también basadas en las TIC. Para conocer la movilidad individual en bicicleta en la ciudad de Madrid, se han recogido los recorridos de los usuarios a través de una aplicación instalada en *smartphones* de ciclistas voluntarios. Con estos recorridos se creó un mapa denominado "*La huella ciclista de Madrid*". Este mapa permite conocer las calles más transitadas, y en consecuencia las zonas que necesitan mayor inversión en infraestructuras ciclistas (Romanillos & Zaltz Austwick, 2016). En Londres se han combinado datos de sistemas de navegación de taxis, información de ocupación de estaciones de bicicletas públicas, y *smartcards* de transporte público para poder realizar clústeres que revelen comportamientos emergentes a nivel individual, y así poder comparar los resultados obtenidos con otros estudios o con los tiempos de tránsito publicados por las propias empresas (Lathia et al., 2013).

2.4.3. Definición de espacios de atracción y generación de viajes

Otros trabajos se han apoyado los datos basados en las TIC para conocer el uso del espacio urbano y, definir espacios de atracción y generación de viajes. Algunos ejemplos son el uso de datos de *Twitter* para cartografiar actividades a partir de distribuciones horarias para extraer espacios de atracción en base a usos del suelo (Frias-Martinez et al., 2012; Hasan, Zhan, & Ukkusuri, 2013), medir la diferencia de densidades de usuarios encontrados respecto al día y la noche (Lansley & Longley, 2016), o proponer delimitaciones alternativas a los espacios administrativos oficiales en función de los espacios de interacción de los usuarios (J. Yin et al., 2017).

Una de las investigaciones de mayor interés en este apartado reside en la identificación de lugares de atracción y generación de lugares en el campo del turismo, donde el desarrollo de las TIC está modificando métodos de investigación y herramientas de administración y marketing (Raun, Ahas, & Tiru, 2016). Las redes sociales son usadas como fuentes para generar contenido como reseñas, opiniones, valoraciones, fotos, etc., con el que se puede estudiar patrones de movilidad, estimar ratios de visita de atracciones específicas, o identificar puntos calientes turísticos en ciudades (Batista e Silva et al., 2018). Gracias a las nuevas fuentes de datos como *Twitter* es posible analizar el comportamiento de los turistas extranjeros, observar los sitios visitados por los turistas y medir la composición turística en el tiempo y en el espacio con mayor precisión (García-Palomares et al., 2015). Las principales fuentes de datos en este caso son redes sociales basadas en fotografías (*Instagram*, *Flickr*, *Panoramio*), plataformas webs de viaje y alojamiento (*AirBnB*, *Booking*, *Tripadvisor*), o *blogs* de viajes y páginas webs (Gutiérrez et al., 2017; Marine-Roig & Anton Clavé, 2015).

Una de las oportunidades más interesantes en la visualización de espacios de atracción y generación de viajes es investigar los espacios de atracción de actividades de diferentes grupos sociales (Longley & Adnan, 2016; Luo, Cao, Mulligan, & Li, 2016; Netto et al., 2015; Shelton et al., 2015). Además, es posible extraer diferencias socioeconómicas en los diferentes comportamientos de movilidad de los turistas al comparar el tipo de red social que utilizan (Salas-Olmedo et al., 2018). En esta línea de trabajo, las *smartcards* también han sido utilizadas para observar el número de residentes con bajo poder adquisitivo que usa el transporte público e identificar trayectos de acuerdo con el nivel socioeconómico (Long & Shen, 2015; Wu et al., 2014).

En este apartado, esta tesis se ha centrado principalmente en la generación de viajes en los campus universitarios como espacios de atracción. Prácticamente todos los estudios de movilidad en el ámbito universitario han utilizado encuestas. Hasta hace pocos años, estas encuestas se han realizado in situ. Tenemos en España ejemplos de estudios realizados a partir de encuestas in situ con el propósito de describir los patrones de movilidad de los estudiantes, trabajadores y residentes en un campus integrado en un entorno urbano (de las Rivas, Iglesias, & Lalana, 2011; Lucas-García, Racero-Moreno, Torrecillas, & García-Sánchez, 2016), o de medir la accesibilidad para poder realizar propuestas de sostenibilidad (Juan M. Albertos, Joan Noguera, María D. Pitarch, 2008; Miralles-Guasch & Domene, 2010; Saladié & Jurado, 2015).

Con el desarrollo de internet, estas encuestas presenciales han pasado a realizarse utilizando páginas webs o correos electrónicos con el objetivo de aumentar las muestras y abaratar costes. Los trabajos basados en encuestas por internet se pueden clasificar en tres categorías. Un primer grupo tiene como objetivo elaborar una descripción sobre las características de transporte de los estudiantes universitarios (Gutiérrez Gallego, Ruiz Labrador, & Rodrigo Muñoz, 2016; Moravec Giormenti et al., 2018; Volosin, Paul, Pendyala, Livshits, & Maneva, 2013; X. Wang, Khattak, Son, & Agnello, 2012; Whalen, Páez, & Carrasco, 2013). Una segunda categoría analiza las características socioeconómicas y de comportamiento social que influyen en la elección del modo de transporte (Delmelle & Delmelle, 2012; Miralles-Guasch, Martínez Melo, & Marquet Sarda, 2014; Seguí-Pons, Ruiz, & Luna, 2013; Soria-Lara, Marquet, & Miralles-Guasch, 2017; J. Zhou, 2014). Una tercera sección busca examinar las emisiones de gases generadas por los viajes diarios a una universidad de estudio (Davison, Ahern, & Hine, 2015; Soria-Lara, Miralles-Guasch, & Marquet, 2017).

En general, las encuestas de todos estos trabajos tienen una estructura similar, con un bloque de preguntas sobre los estudiantes y sus características socioeconómicas y otro sobre los hábitos de movilidad diaria y modos de viaje empleados. Habitualmente, el campo de estudio es una única universidad. Todas estas encuestas presentan un periodo temporal no superior a un mes. De forma general, del 10 al 30% de las encuestas enviadas fueron contestadas. En algunos casos se han combinado las encuestas con otros métodos como entrevistas (Miralles-Guasch et al., 2014). Las muestras son mayores respecto a las encuestas realizadas in situ, pero siguen presentando un bajo detalle temporal y dificultad de actualización de los datos. El uso de fuentes basadas en *Big Data* para el estudio de la

movilidad universitaria ha empezado a entrar muy recientemente, por lo que apenas hay constancia de trabajos que hayan empleado nuevas fuentes de datos en combinación con datos de encuestas (Delclòs-Alió & Miralles-Guasch, 2017).

2.4.4. Impactos de eventos en la ciudad

El impacto de los eventos de masas es un tema que presenta una interesante oportunidad para su análisis a partir de datos de las TIC en general, y de las redes sociales como *Twitter* en particular. Tradicionalmente se han empleado cuestionarios y encuestas sobre la población turística en eventos como el Mundial de fútbol de 2010 celebrado en Sudáfrica (Knott, Swart, & Visser, 2015). Sin embargo, en los últimos años, ha aumentado el interés por detectar el impacto de eventos a partir de las nuevas fuentes de datos, ya que no solo permiten identificar puntos de interés, sino también pueden sacar datos sobre la ubicación o el evento (H. Li, Ji, & Zhao, 2015). Así, por ejemplo, encontramos el uso de *Bluetooth* para estudiar el comportamiento y la densidad de la población en los festivales musicales de Gante (Bélgica) (Versichele, Neutens, Delafontaine, & Van de Weghe, 2012); el empleo varias fuentes de *Big Data* (*Airbnb*, *Waze*, y CDRs de telefonía móvil) para estimar el impacto de los megaeventos en el tráfico de la ciudad de Río de Janeiro durante los Juegos Olímpicos de 2016 (Xu & González, 2017), o el uso de tarjetas inteligentes de transporte para la predicción de comportamientos del uso del transporte público bajo eventos especiales celebrados en Singapur (Pereira, Rodrigues, & Ben-Akiva, 2015). En España, *Vodafone* y *CARTO* estudiaron la *MTV Music Week Bizkaia* combinando una aplicación móvil desarrollada por ellos y sensores ubicados en el evento[6]. En el caso concreto de los estudios del impacto de manifestaciones LGBT, hay una primera aproximación realizada por BBVA analizando la semana del desfile nacional del Orgullo de Madrid de los años 2011 y 2012. Usando los datos de transacciones realizadas con las tarjetas bancarias, se detectó un aumento de gasto del 24% en la ciudad respecto a semanas anteriores o posteriores al evento. También observaron una tendencia creciente del 9% en el consumo de 2012 respecto al año 2011, y una distribución espacial importante sobre todo en el distrito Centro[7].

[6] http://www.saladeprensa.vodafone.es/c/notas-prensa/np_soluciones_mtvmusicweek/
[7] https://issuu.com/cibbva/docs/big_data_english

La detección de eventos mediante datos de *Twitter* ha sido un tema recurrente en investigaciones de la última década. Estas investigaciones se han centrado en el desarrollo de metodologías con el objetivo de crear modelos de predicción y detección de eventos. Estos modelos usan datos geográficos o el propio texto de los *tweets* a partir de técnicas de *datamining*, y los asocian a un lugar determinado. En esta línea, (R. Lee & Sumiya, 2010) diseñaron un sistema de detección e identificación de eventos en tiempo real para identificar festivales locales en Japón. (Weng, Yao, Leonardi, & Lee, 2011) propusieron un método teórico de detección de eventos basado en la agrupación a partir de palabras. (Abdelhaq, Sengstock, & Gertz, 2013) elaboraron un sistema de detección de eventos a partir de la extracción de mensajes con determinadas palabras claves en un determinado ámbito temporal, su identificación en un área espacial, y la formación de grupos para su análisis. (Popescu & Pennacchiotti, 2011) clasificaron los eventos relacionados con famosos en diferentes categorías, estudiaron la reacción de las personas mediante sus *tweets* a esos eventos mediante análisis de sentimientos, y realizaron índices de intensidad y controversia. (Kim, Kojima, & Ogawa, 2016) propusieron un sistema de identificación de eventos en Estados Unidos a partir de la captura de palabras claves que permitan la identificación de comunidades locales. (Y. Huang, Li, & Shan, 2018) se basaron en el análisis de texto de los mensajes mediante el algoritmo LDA para detectar eventos pequeños y determinar sus patrones espacio-temporales. (Zhañay, Cordero, Cordero, & Urigüen, 2019) diseñaron un modelo en Cuenca (Ecuador) que detectase eventos a partir de los textos de los *tweets* y los vinculase a un lugar, obteniendo valores de precisión sobre el 70%.

Un campo de investigación de gran interés es la extracción de patrones espaciales o espacio-temporales a partir de datos de *Twitter* para analizar los sentimientos en diferentes eventos. (Chin, Zappone, & Zhao, 2016) analizaron a nivel de estado los sentimientos encontrados en los *tweets* respecto a cada candidato de las elecciones presidenciales de Estados Unidos de 2016. (Conover et al., 2013) estudiaron la distribución geográfica y los flujos de comunicación interestatales de un movimiento social como las protestas que se celebraron en Estados Unidos en 2013 bajo el nombre *Occupy Wall Street*. Los eventos de ámbito deportivo han ocupado un papel importante en este tipo de estudios.(Kirilenko & Stepchenkova, 2017) estudiaron la distribución de *tweets* tanto en Rusia como a nivel global que comentaban los Juegos Olímpicos de Sochi en el año 2014. (Kovacs-Gyori, Ristea, Havas, Resch, & Cabrera-Barona, 2018)

analizaron las distribuciones espacio-temporales de tanto los turistas como los residentes de la ciudad de Londres durante los Juegos Olímpicos de 2012. (Steiger, de Albuquerque, et al., 2015) trataron el impacto de las series mundiales de béisbol en la ciudad de Boston a partir de clústeres de *tweets* próximos tanto espacio-temporalmente como semánticamente.

Otros eventos de interés para la investigación son los de causas naturales para su prevención y tratamiento de posibles desastres. Entre estos fenómenos naturales destacan los huracanes (Hiltz et al., 2014) y los terremotos (Sakaki, Okazaki, & Matsuo, 2010). También es posible comparar eventos de diversas índoles como un evento político-social con un evento natural. (X. Zhou & Xu, 2017) compararon la evolución a lo largo de las horas del día de un evento natural (tormenta torrencial) y un evento social (visita del Papa Francisco) en la región Nueva York – Washington DC.

2.4.5. Información para conocer la percepción del transporte

El análisis semántico y de sentimientos de los textos publicados por los usuarios de *Twitter* es un campo prolífico con una amplia variedad de estudios en esta última década, tanto en el aspecto metodológico como en el analítico, gracias a la facilidad de recolectar muestras de datos con riqueza en sentimientos y opiniones en poco tiempo. La metodología en estos trabajos es normalmente similar, con un foco principal en técnicas de minería de texto para extraer temas y palabras, y análisis de sentimientos para añadir un valor positivo, negativo, o neutral a los textos a partir de las palabras que contienen. Hay una importante cantidad de investigaciones que usa el modelo *Latent Dirichlet Allocation* (LDA) para clasificar los *tweets* en grupos relacionados con el número de veces que una palabra es escrita y su relación con otras palabras.

En el campo de la geografía, la minería de texto para la extracción de temas y sentimientos ha sido principalmente usado para mapear y comparar la frecuencia sentimientos durante el día a diferente escalas (Biever, 2010; Kocich, 2017; Lansley & Longley, 2016; Mitchell, Frank, Harris, Dodds, & Danforth, 2013; Steiger, Resch, & Zipf, 2016; Wachowicz & Liu, 2016). Otros campos relevantes son las percepciones y sentimientos del uso de espacios verdes en la ciudad en diferentes momentos del día (Kovacs-Györi et al., 2018; Lim et al., 2018), la localización de sentimientos para resolver problemas

basados con la salud como la obesidad (Ghosh & Guha, 2013), o el análisis de eventos (ya visto en el apartado anterior).

Todavía hay pocos trabajos que han usado el texto de *Twitter* como fuente de datos para el análisis de datos en el transporte público, aunque es un campo que está en auge en los últimos años. (Collins, Hasan, & Ukkusuri, 2013) analizó la distribución temporal de sentimientos de los *tweets* de los usuarios del Sistema de transporte público *Chicago Transit Authority* en la ciudad de Chicago. Observaron que sobre el 25% de los datos recogidos son relevantes y basados en opiniones, principalmente con sentimientos negativos sobre el servicio (Schweitzer, 2014) recogió *tweets* que contenían la palabra *SEPTA* (acrónimo de la agencia de transporte público de la región de Filadelfia de Estados Unidos, y compararon la distribución de *tweets* relacionados con el transporte con otros temas. (Luong & Houston, 2015), desarrollaron un análisis de sentimiento sobre las líneas de trenes urbanos de la ciudad de Los Ángeles usando *tweets* localizados en un radio de 50 millas de la ciudad, seleccionando los usuarios que interactuaron con las cuentas de *Twitter* de las diferentes líneas de tren. (S. Zhang & Feick, 2016) clasificaron por relevancia *tweets* de la región de Waterloo (Canadá) durante un periodo de 16 meses, hallando que el 99% de los *tweets* relacionados con el transporte público no trataban en realidad el tema.

Centrándonos en trabajos de los últimos tres años, (Casas & Delmelle, 2017) descargaron *tweets* usando palabras claves relacionadas con el servicio de transporte público metropolitano de Cali (Colombia), y extrajeron paradas de autobús, rutas, y temas principales a partir de los textos de los *tweets*. Encontraron que la seguridad era el tema que más preocupaba a los usuarios del servicio. (Kulkarni, Abellera, & Panangadan, 2018) crearon clústeres de temas de *tweets* que comentaban el transporte público en California sacando como observaciones que la calidad de la identificación de temas depende del tamaño de la muestra, el número de temas a especificar, y el valor semántico. (Haghighi et al., 2018) evaluó las opiniones de los usuarios del transporte público de Salt Lake City (EEUU) tras extraer temas latentes de los *tweets* y realizar un análisis de sentimiento por día y tipo de transporte. (Hosseini, El-Diraby, & Shalaby, 2018) analizó datos de *Twitter* de tres agencias de transporte de Canadá (dos ubicadas en Toronto, y otra en Vancouver). A partir de extracción de temas y comparación de su distribución, encontraron que los tres temas principales en los tres servicios eran la seguridad, los tiempos de viaje, y el nivel de servicio de las agencias. (El-Diraby, Shalaby, & Hosseini,

2019) continuó este trabajo, añadiendo un análisis de sentimientos que reveló una mayor satisfacción de los usuarios en fines de semana.

Por último, están empezando a aparecer trabajos enfocados en la creación de modelos para detectar problemas concretos en una red de transporte público a partir de la información semántica de los datos de *Twitter*. Como ejemplo, (Ji et al., 2018) se concentraron en crear un modelo que detectase similitudes semánticas en quejas escritas en diferentes *tweets* localizados en un mismo lugar. Este modelo está basado en la conectividad especial entre líneas de Metro y un vocabulario común de quejas. El caso de estudio para probar el modelo fue el servicio de transporte público metropolitano de Washington.

2.5. Aportación de la investigación

La aportación principal de la investigación realizada en esta tesis doctoral consiste en analizar en profundidad el valor de *Twitter* como nueva fuente de datos basada en las TIC para el estudio de la distribución de la población metropolitana y sus patrones de movilidad. Se utilizan diferentes casos de estudio vinculados a las distintas aplicaciones de las nuevas fuentes de datos para la investigación de la movilidad metropolitana desarrollados en el apartado anterior. Esta tesis busca exprimir las diferentes capas de información de un *tweet* (información espacial, información temporal e información semántica) y testear la validez de cada una de estas dimensiones contrastando los resultados obtenidos con otras fuentes de datos oficiales.

Para analizar comportamientos de movilidad general, esta tesis propone utilizar los datos de *Twitter* para construir matrices OD. Para ello, se ha buscado mejorar el método utilizado para construir matrices OD con datos de telefonía, con el objetivo de mejorar y validar *Twitter* como fuente alternativa frente a los datos de telefonía. Sin embargo, salvo Lee et al. (2015), ninguno de los trabajos mencionados en el apartado 2.4.1 de la tesis ha cruzado datos de *Twitter* con datos de usos del suelo para mejorar la definición de la residencia y el lugar de trabajo. En esta tesis, si se realizó este proceso, enriqueciendo los datos de *Twitter* con una capa de alto detalle espacial de usos del suelo del Catastro. A continuación, se completaron todos los pasos para obtener las matrices de viajes, expandiendo los datos (basándose en dos diferentes fuentes: censo de población y

registros de la Seguridad Social) y validando los resultados con dos escalas de agregación espacial.

Con el objetivo de visualizar la movilidad espacio-temporal de los usuarios de *Twitter*, esta tesis aprovecha la facilidad de transformar los *tweets* en entidades de puntos con los que poder diseñar caminos espacio-temporales en tres dimensiones. Mientras que ha habido trabajos previos que construyeron caminos espacio-temporales o que usaron *Twitter* como fuente de datos para analizar las características espacio-temporales de la movilidad individual, apenas hay estudios que han empleado *Twitter* para la construcción de caminos espacio-temporales. Este trabajo ha querido combinar estas dos vertientes. Sin embargo, los datos de *Twitter* capturan tiempo y localización, pero no recogen información más detallada como la naturaleza de los eventos o actividades realizadas (Q. Huang & Wong, 2015). Por ello, una mejora metodológica que realiza este trabajo para paliar esta desventaja es la combinación del mapeado de caminos espacio-temporales creados a partir de datos de *Twitter* con datos de usos del suelo. De los trabajos mencionados en el apartado 2.4.2. de la tesis que emplean datos de *Twitter*, solo se consta un trabajo que haya combinado estos datos con la información de usos del suelo (García-Palomares et al., 2018).

En cuanto al análisis de la movilidad universitaria, la originalidad del presente trabajo respecto a estudios anteriores comentados en el apartado 2.4.3. de la tesis se halla en el filtrado de *tweets* publicados en facultades en horario docente para identificar potencial población universitaria, los campus a los que asisten y sus lugares de residencia. Este método permite obtener de forma rápida y barata una muestra mayor que las obtenidas por encuestas tradicionales. La utilidad de esta investigación radica en la obtención de información complementaria que permita estimar las residencias de la población universitaria, información que existe ya que las universidades la obtienen cuando los estudiantes realizan la matrícula, pero que con frecuencia no es fácil acceder. Otras innovaciones consisten en la visualización de áreas de influencia de las distintas universidades y en el diseño de un modelo gravitacional de *Huff* para comparar los datos obtenidos de *Twitter* con la situación que se da en la realidad.

La originalidad del análisis de los patrones espacio-temporales y la distribución de usuarios durante un megaevento celebrado en una ciudad radica en el uso de datos geolocalizados de *Twitter* para investigar en profundidad el impacto del evento. Mientras que ha habido varios trabajos comentados en el apartado 2.4.4. de la tesis y basados en la

detección de eventos a partir del contenido geográfico o semántico de los datos, hay muy pocos estudios que tratan la huella espacio-temporal de eventos a partir de datos de redes sociales. Para ello se propone el uso de técnicas geoestadísticas, como el análisis de autocorrelación espacial o análisis de clústeres, para ahondar en la impronta espacial del evento, en contraposición con otros trabajos que han empleado *Twitter*, pero se han limitado a visualizar el conteo de *tweets*. También se emplean técnicas de cartografía temporal para poder visualizar los cambios de la distribución de la población en diferentes momentos del evento. Además, este estudio otorga una metodología de identificación de la provincia o país de origen de cada visitante a partir de la descarga de los 3.200 últimos mensajes de cada usuario detectado en el caso de estudio, un aspecto que no se ha tratado en trabajos previos y que multiplica el valor de la aportación de este trabajo. Al utilizar identificadores numéricos y trabajar en una escala de provincias y países se mitigan los problemas de privacidad que conlleva el uso de datos de *Twitter*.

Para la investigación de la percepción que tienen los usuarios de *Twitter* sobre un sistema de transporte público, esta tesis se ha centrado en extraer temas y sentimientos y en representar su distribución en el área de estudio. Aunque casi todas las investigaciones mencionadas en el apartado 2.4.5. de la tesis tienen en común una metodología similar de extracción de temas y análisis de sentimientos de los *tweets* relacionados con servicios de transporte público, pocos trabajos han trabajado propiamente en la dimensión espacial (estando limitada normalmente a la detección de eventos), o han validado los resultados con datos oficiales. Además, apenas hay trabajos que investiguen los factores geográficos que puedan explicar la distribución espacial de los usuarios o temas de las quejas relacionados con un sistema de transporte público. Esta tesis plantea ir más allá, cartografiando la distribución de usuarios y temas en el área de estudio, y añadiendo un método para analizar la causalidad espacial detrás de la distribución de los usuarios en la red de transporte. Para este objetivo, se propone un modelo de Regresión Geográficamente Ponderada (GWR) para cartografiar y evaluar los factores que contribuyen a la distribución espacial de los usuarios con sentimientos negativos en el área de estudio.

3. ÁREA DE ESTUDIO, DATOS Y METODOLOGÍA

3.1. Área de estudio

3.1.1. El Área Metropolitana de Madrid

El Área Metropolitana de Madrid está situada en el centro de la Península Ibérica, y se trata del área metropolitana de mayor población, actividad y servicios de España. Tiene una población estimada en más de 6 millones de habitantes en el año 2020[8]. Podemos decir que uno de cada ocho habitantes de España (casi el 13% de la población) reside en el Área Metropolitana de Madrid. Además, es la cuarta urbe más poblada de la Unión Europea, solo por detrás de París, Londres, y la Región del Ruhr. Según la *Globalization and World Cities Research Network* (GaWC), Madrid es un área metropolitana *alpha*, es decir, un área metropolitana muy importante a escala global que enlaza regiones económicas de gran importancia en la economía mundial[9]. Más de 3.1 millones de personas trabajan en el área de estudio (según datos del Mercado de Trabajo del año 2019 del Instituto de Estadística de la Comunidad de Madrid)[10].

Debido a la falta de ordenación legal del área metropolitana de Madrid es difícil definir que municipios conforman dicha área metropolitana. En esta tesis doctoral, se ha escogido la delimitación del Instituto de Estadística de la Comunidad de Madrid como fuente para definir el Área Metropolitana de Madrid. Esta configuración está formada por 50 municipios. El municipio de Madrid a su vez está dividido en 21 distritos, y cuenta con más de 3.2 millones de habitantes, englobando más del 53% de la población y más del 67% de la bolsa de trabajadores del área metropolitana.

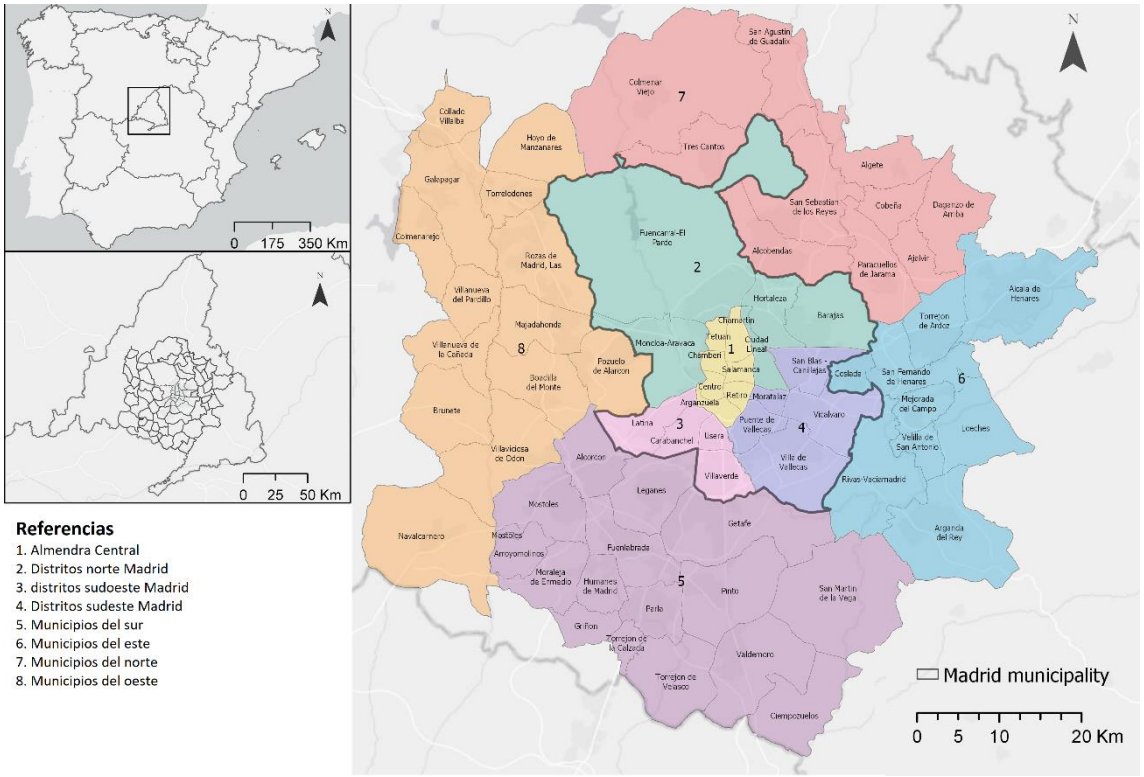
Debido a la centralización de servicios y al desarrollo de las infraestructuras de transporte, en las últimas décadas los municipios periféricos han desarrollado un fuerte aumento de la población, convirtiéndose en ciudades dormitorio (García-Palomares & Gutiérrez-Puebla, 2007). Destaca principalmente el cinturón de cinco municipios del sur del área metropolitana formado por Getafe, Leganés, Alcorcón, Fuenlabrada y Móstoles (cada ciudad cuenta con más de 150.000 habitantes). También se están formando ciudades dormitorios en el norte del área metropolitana (principalmente en los municipios de Alcobendas y San Sebastián de los Reyes) y en el este (destacan Coslada, San Fernando de Henares y Torrejón de Ardoz, municipios ubicados en el eje Madrid – Alcalá de

[8] http://www.madrid.org/iestadis/fijas/estructu/demograficas/padron/estructupopc_prov.htm
[9] https://www.lboro.ac.uk/gawc/world2018t.html
[10] http://www.madrid.org/iestadis/fijas/estructu/sociales/iss19.htm

Henares). En este eje este, que destacar la ciudad de Alcalá de Henares, ya que, a pesar de integrar el área metropolitana, ha tenido históricamente un peso propio de población al estar más alejada de Madrid que otras ciudades dormitorio, pero ser la primera ciudad de paso en el eje Madrid – Zaragoza – Barcelona. En los últimos años está empezando a haber también un crecimiento en la parte oeste del área de estudio (principalmente en Las Rozas de Madrid, Pozuelo de Alarcón y Boadilla del Monte). En contraposición con los municipios poblados principalmente por gente de clase obrera en el este o norte de la ciudad, los municipios del oeste metropolitano destacan por un alto nivel de renta.

La Figura 14 recoge los municipios (y distritos de la ciudad de Madrid) del Área Metropolitana de Madrid, y los agrega en ocho grandes zonas metropolitanas, con niveles similares de población y comportamientos propios de movilidad. La tabla 4 recoge la suma de la población total y el número de trabajadores además del nivel de renta medio de cada zona metropolitana.

Figura 14: Mapa del Área Metropolitana de Madrid.



Fuente: Elaboración propia.

Tabla 4: Distribución de la población por zonas metropolitanas.

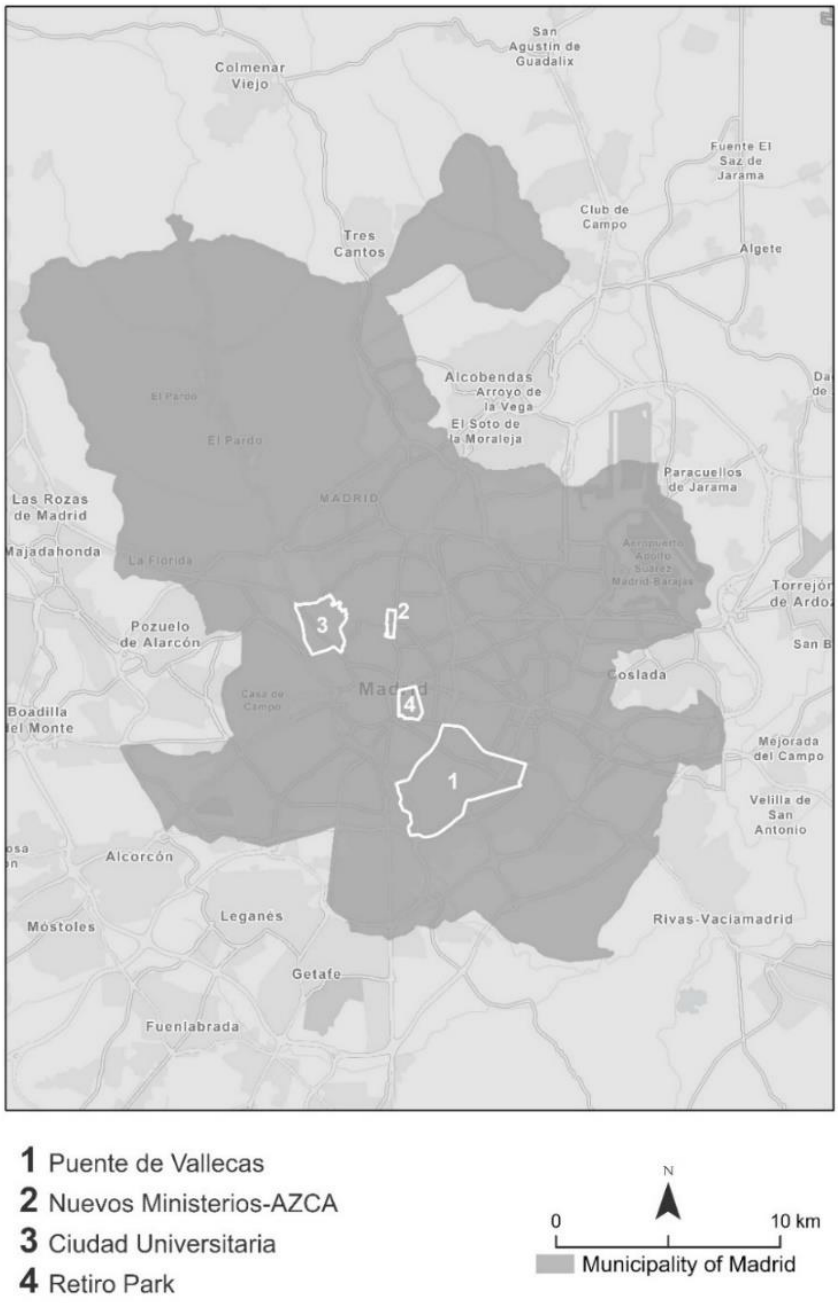
Zonas	**Residentes**	**Trabajadores**	**Renta media**
Almendra Central	983.625	995.418	32.601,17
Distritos norte	830.307	486.648	30.191,92
Distritos Suroeste	734.981	184.728	16.130,38
Distritos Sureste	656.477	426.052	18.589,64
Sur metropolitano	1.295.725	348.601	17.199,36
Este metropolitano	627.997	202.823	18.052,58
Norte metropolitano	370.037	225.288	25.550,64
Oeste metropolitano	535.176	263.771	35.663,78
Total	**6.034.325**	**3.133.329**	**24.247,43**

Fuente: Elaboración propia a partir de datos del Instituto de Estadística de la Comunidad de Madrid y del Portal de Datos Abiertos del Ayuntamiento de Madrid.

Dentro del municipio de Madrid se han escogido cuatro zonas especializadas para la elaboración del caso de estudio correspondiente al apartado 4.2. de la tesis doctoral. Estas zonas cuentan con actividades y usos del suelo dominantes o específicos. El objetivo a la hora de escoger dichas zonas es recoger la actividad de los usuarios de *Twitter* en días laborables a cada hora, y usar estos resultados como proxy para estudiar la dinámica en la utilización de los usos del suelo del Área Metropolitana de Madrid a lo largo del día (Figura 15).

1. Distrito de Puente de Vallecas. – Este distrito es un espacio netamente residencial y una de las zonas más densamente pobladas de la ciudad de Madrid. Para recoger usuarios de *Twitter* asociados a este espacio residencial, se han seleccionado los usuarios que han *twitteado* habitualmente en este distrito en horario de noche (8 a 21 horas), de manera que podemos pensar son usuarios residentes en dicho espacio.

Figura 15: Zonas de estudio a analizar dentro del municipio de Madrid.



Fuente: Elaboración propia.

2. Complejo Nuevos Ministerios-AZCA. – Situado en el eje de la Castellana (una de las arterias principales de la ciudad que conecta el centro con el sector norte), es una de las principales zonas empresariales y financieras de Madrid, con un importante número de empleos. En este caso, se han seleccionado usuarios de *Twitter* que han *twitteado* habitualmente en esta zona en horario de mañana (8 a 15 horas), considerando por tanto que se trata de usuarios cuyo empleo está en esta zona.

3. Ciudad Universitaria. – Es el principal campus universitario de la ciudad, donde se concentran las principales facultades de la Universidad Complutense de Madrid (UCM), pero también de la Universidad Politécnica de Madrid (UPM), la Universidad Nacional a Distancia (UNED) y otras universidades de ámbito privado. Nuevamente la selección de usuarios de *Twitter* se ha realizado a partir de aquellos usuarios que han *twitteado* en este espacio en horario de mañana (8 a 15 horas).

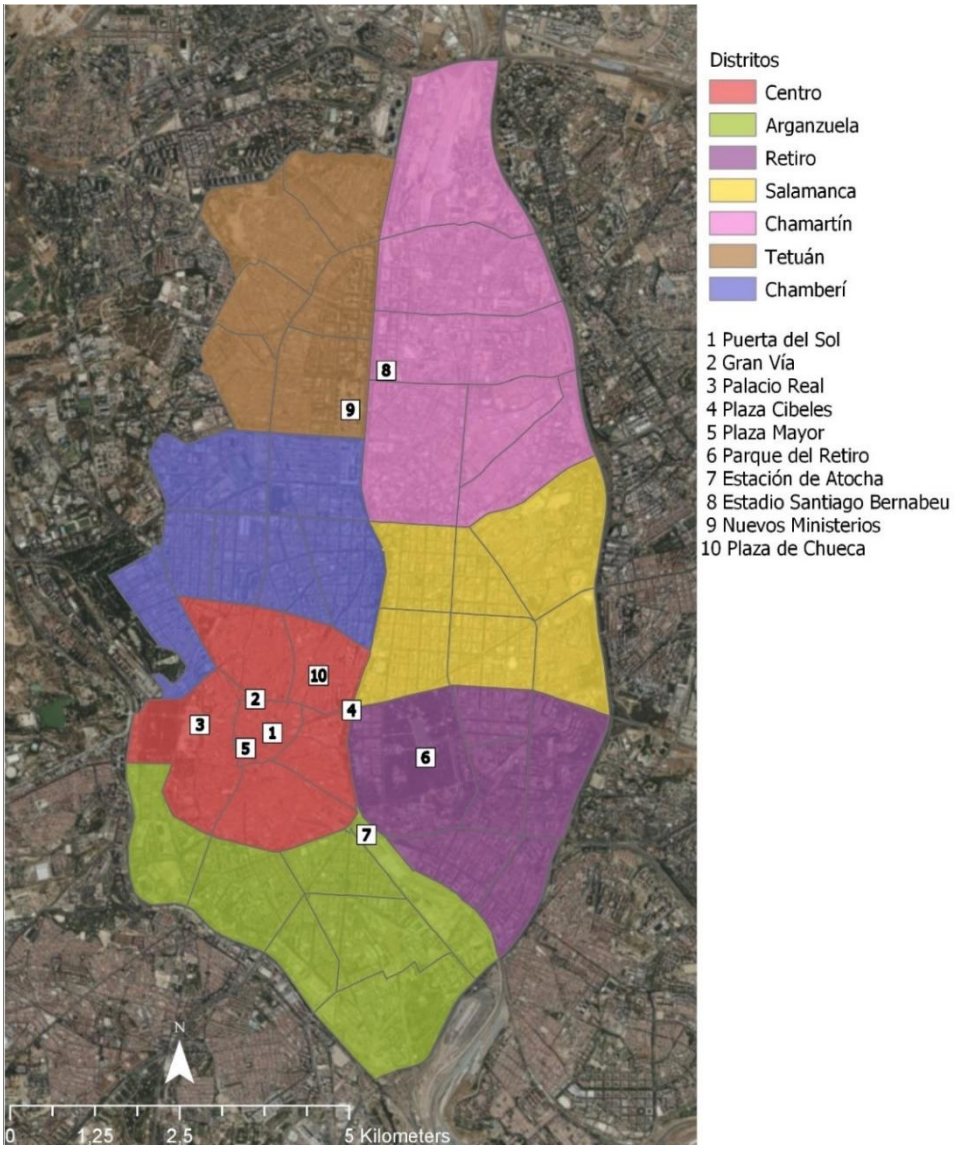
4. Parque del Retiro. – Uno de los principales espacios verdes y atractivo turístico del centro de la ciudad, es una de las zonas de ocio más utilizadas y uno de los pulmones más importantes de Madrid. En este caso, los usuarios de *Twitter* asociados al parque son aquellos que han *twitteado* habitualmente en horario de tarde y tarde-noche (16 a 19 horas).

3.1.2. La Almendra Central de Madrid

Dentro del municipio de Madrid, destaca la Almendra Central, con una población estimada en casi 1 millón de habitantes, más del 30% de la población del municipio. Esta zona está comprendida por los siete distritos dentro de la carretera de circunvalación M-30 (Figura 16). Se trata del área de la ciudad de mayor concentración de actividad y servicios, y donde se halla la mayor parte del empleo de la ciudad (la Almendra Central cuenta con casi el 50% de trabajadores del municipio de Madrid).

El distrito de la Almendra Central que genera un mayor impacto es el Distrito Centro, corazón y principal foco turístico del municipio. Este distrito es el espacio de mayor actividad de todo el municipio, englobando los principales lugares de interés de la ciudad (Puerta del Sol, Plaza Mayor, Gran Vía, Palacio Real, etc.), y barrios con una identidad propia y conocida (barrios de Malasaña, Chueca, Lavapiés, o Letras) que atraen tanto a la población joven autóctona como a visitantes y turistas. La actividad en el Distrito Centro es tan grande que a finales de 2018 el Ayuntamiento de Madrid convirtió el distrito en una zona de bajas emisiones (*Madrid Central*) con el objetivo de limitar las emisiones de tráfico privado. Otros distritos de la Almendra Central también a destacar son Arganzuela (distrito donde se ubica la estación de trenes Madrid-Atocha, la principal estación de tren de la ciudad y del país) y Chamartín (distrito financiero y económico de la ciudad).

Figura 16: Distritos y barrios de la Almendra Central de Madrid.



Fuente: Elaboración propia.

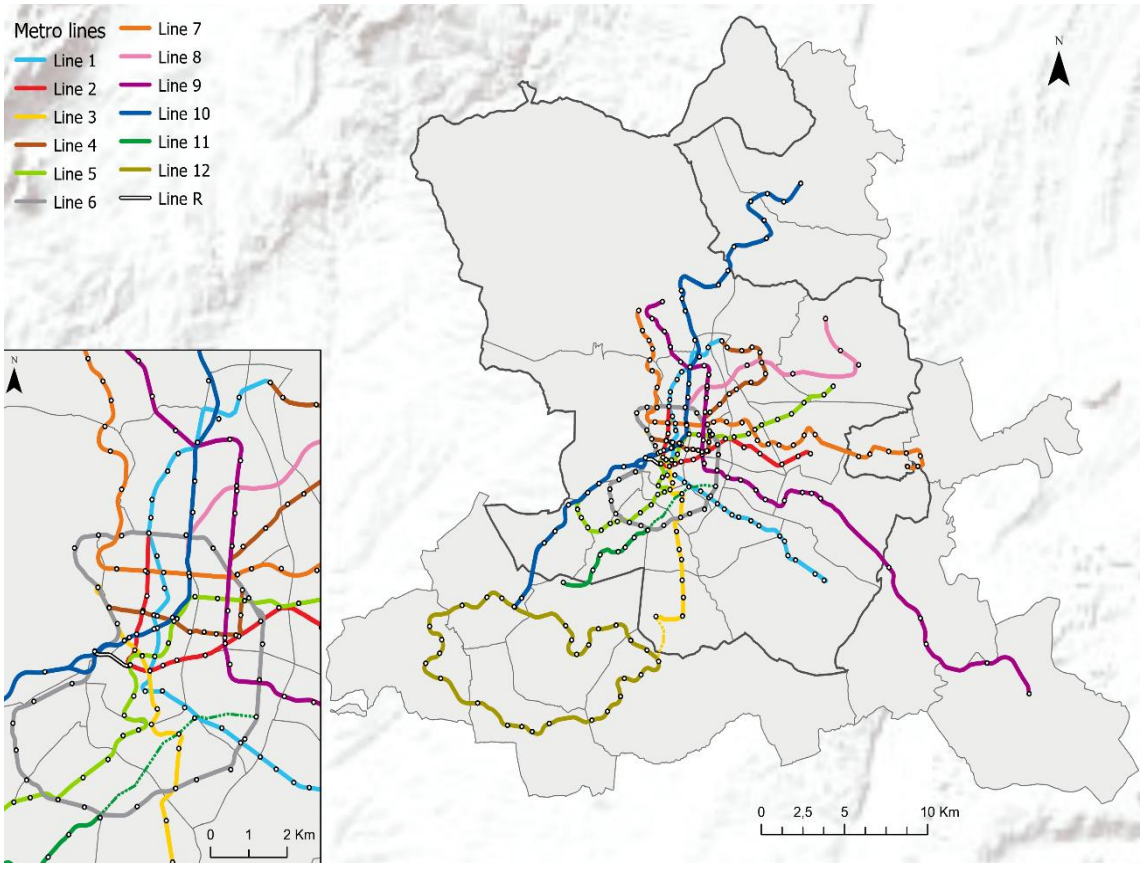
Fuera de la Almendra Central, cabe señalar los distritos del sur del municipio de Madrid (Carabanchel, Latina, Puente de Vallecas) por ser las principales zonas dormitorio de la ciudad; los distritos del este (Ciudad Lineal, Hortaleza), zonas de residencia de población con alto nivel de renta; el distrito de Moncloa-Aravaca, sede del principal campus universitario del área metropolitana; y el distrito de Barajas, en este caso por albergar el aeropuerto de Madrid.

3.1.3. El Metro de Madrid

El Área Metropolitana de Madrid cuenta con una variedad de servicios de transporte público integrados por una flota de autobuses urbanos e interurbanos, una red de trenes de Cercanías que conecta una cantidad importante de municipios del área metropolitana con el municipio central, y una red de Metro que da servicio a la ciudad de Madrid, pero también a los municipios colindantes. Todos estos servicios de transporte públicos están gestionados por el Consorcio Regional de Transportes de Madrid. Esta institución ofrece diversos billetes y tarifas para el uso de sus servicios, destacando los abonos de transporte. Estos abonos son tarjetas inteligentes válidas para cualquier servicio de transporte de la región y ofrecen un número ilimitado de viajes durante su periodo de validez.

El Metro de Madrid está formado por una red de casi 300 kilómetros de longitud que está integrada por 12 líneas convencionales, una línea ramal, y más de 300 estaciones. La red de Metro se extiende a todos los distritos del municipio de Madrid y alcanza un total de 14 municipios (Figura 17, Tabla 5).

Figura 17: Red de Metro de Madrid en el Área Metropolitana.



Fuente: Elaboración propia.

Tabla 5: Estructura de las líneas convencionales de Metro de Madrid.

Línea	**Estaciones**	**Longitud**	**Frecuencia**	**Número viajeros en 2018**
1	33	23,320 km	4 min	95.549.987
2	20	14,031 km	4 min	43.969.307
3	18	16,424 km	3 min	66.538.578
4	23	14,625 km	3,5 min	43.442.442
5	32	23,207 km	4 min	69.848.412
6	28	23,472 km	4 min	107.544.619
7	31	32,919 km	5,5 min	44.252.587
8	8	16,459 km	4,5 min	18.928.919
9	29	39,500 km	5,5 min	43.415.474
10	31	36,514 km	5 min	75.130.369
11	7	8,500 km	6 min	5.421.189
12	28	40,596 km	6,5 min	32.109.243

Fuente: Elaboración propia a partir de datos del Consorcio de Transportes de Madrid.

El Metro de Madrid es el servicio de transporte público más utilizado en el municipio de Madrid. Este sistema de transporte tiene un uso estimado de 2,3 millones de viajeros por día[11]. Cerca del 43% de los ciudadanos emplean el Metro de Madrid para viajar, mientras que el 27% viaja en autobús, y solo el 13% de los viajeros utilizan el transporte de Cercanías. El motivo principal de desplazamiento por los usuarios del Metro son los viajes de residencia al lugar de trabajo (55,03%), o de estudio (15,50%) [12].

Cabe mencionar el servicio de Cercanías, propiedad de la empresa ferroviaria *RENFE*, que cuenta con 370 kilómetros, 10 líneas y 90 estaciones. Este servicio opera a una escala geográfica mayor que la red de metro, y alberga a todo el Área Metropolitana de Madrid. Cercanías es el principal medio de transporte para los ciudadanos de municipios más alejados del municipio de Madrid. No todos los municipios del Área Metropolitana de Madrid cuentan con servicio de Cercanías. Los residentes de estos municipios emplean el servicio de autobuses interurbanos para desplazarse al municipio central.

[11] https://www.metromadrid.es/es/quienes-somos/metro-de-madrid-en-cifras
[12]https://www.metromadrid.es/sites/default/files/documentos/Portal%20de%20transparencia/Memorias/INFORME%20CORPORATIVO%202018.pdf

3.1.4. Las universidades del Área Metropolitana de Madrid

El Área Metropolitana de Madrid alberga seis universidades públicas y siete universidades privadas. En el curso 2017/18 las universidades madrileñas contaron con más de 315.000 estudiantes inscritos, de los cuales aproximadamente más de 210.000 se encontraban matriculados en universidades públicas o centros adscritos, unos 65.000 en universidades privadas, y casi unos 40.000 restantes en universidades especializadas en estudios a distancia[13]. La universidad pública más importante es la Universidad Complutense de Madrid, la cual concentra más del 25% de la oferta universitaria del Área Metropolitana de Madrid (Tabla 6).

Tabla 6: Universidades del Área Metropolitana de Madrid.

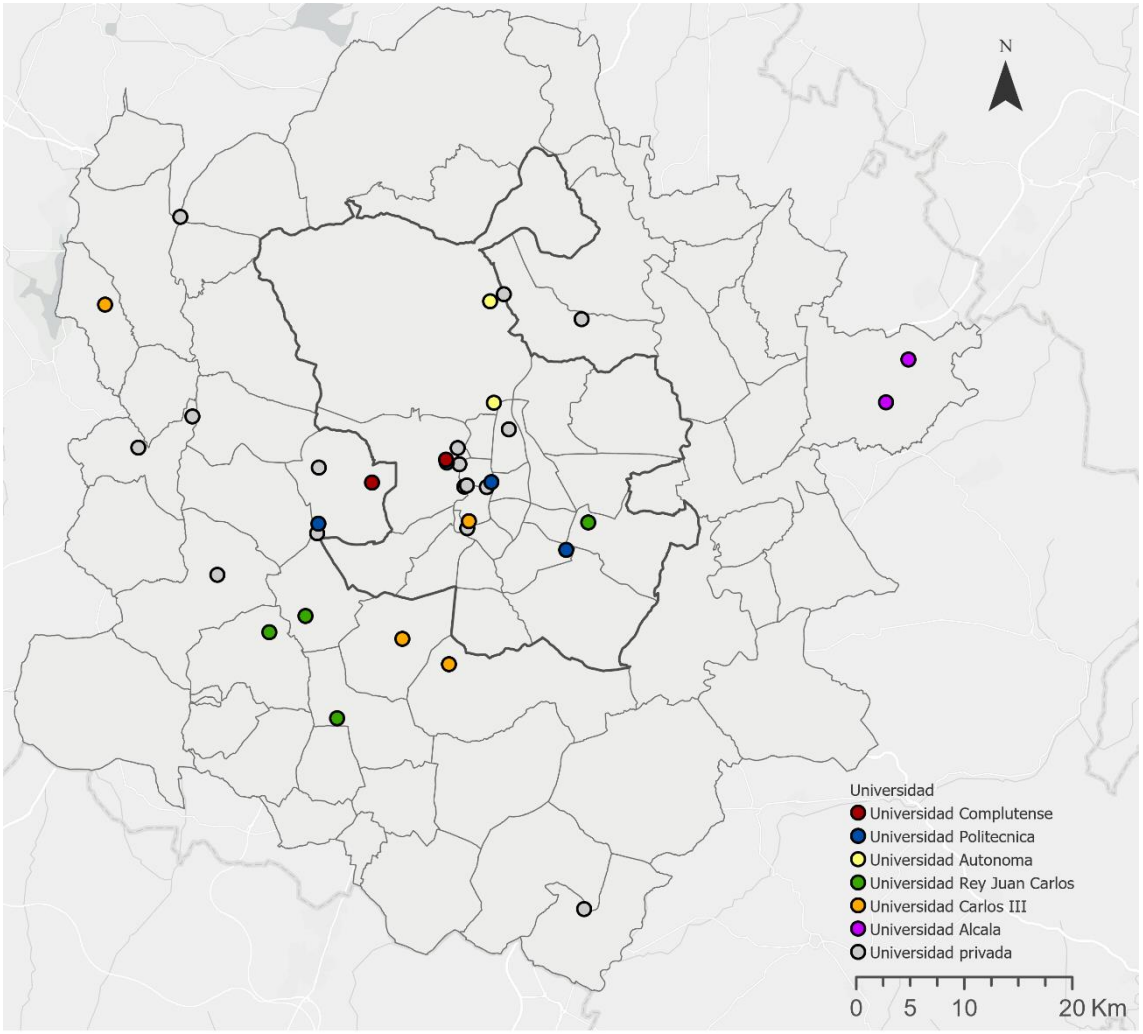
Universidad	**Tipo**	**Número de campus**	**Número de estudiantes en 2017**
Universidad Complutense de Madrid	Pública	2	72.738
Universidad Politécnica de Madrid	Pública	4	36.694
Universidad Autónoma de Madrid	Pública	2	47.476
Universidad Rey Juan Carlos	Pública	4	29.285
Universidad Carlos III	Pública	4	19.232
Universidad de Alcalá	Pública	2	17.983
Universidad Europea de Madrid	Privada	2	11.421
Universidad CEU San Pablo	Privada	2	8.640
Universidad Francisco de Vitoria	Privada	1	8.206
Universidad Pontificia de Comillas	Privada	3	9.142
Universidad Camilo José Cela	Privada	2	9.916
Universidad Antonio de Nebrija	Privada	3	6.057
Universidad Alfonso X el Sabio	Privada	2	6.854

Fuente: Elaboración propia a partir de datos del Ministerio de Educación, Cultura y Deporte.

[13] http://www.comunidad.madrid/servicios/educacion/sistema-universitario-madrileno

La Figura 18 muestra la ubicación de los campus universitarios en el Área Metropolitana de Madrid. Tanto la Universidad Complutense como la Politécnica de Madrid están orientadas a ofrecer servicios a estudiantes de toda la comunidad de Madrid. La Universidad Autónoma de Madrid también tiene esa orientación, pero por su ubicación al norte del municipio recibe principalmente estudiantes del norte del Área Metropolitana. La Universidad de Alcalá, fundada en 1499, fue durante siglos la universidad vinculada con Madrid hasta la fundación de la Universidad Complutense de Madrid en la capital. Por último, el fuerte crecimiento de la población en los municipios del sur del Área Metropolitana de Madrid ha conllevado la creación de la Universidad Rey Juan Carlos (con rectorado en Móstoles) y la Universidad Carlos III (con rectorado en Getafe), ambas nacidas en las últimas décadas del siglo XX.

Figura 18: Ubicación de los campus universitarios en el Área Metropolitana de Madrid.



Fuente: Elaboración propia.

Casi 80.000 estudiantes están adscritos a los campus ubicados en Ciudad Universitaria (ya señalada en la Figura 16). Es decir, Ciudad Universitaria atrae a más de un 30% de la oferta universitaria del Área Metropolitana de Madrid. Además de facultades y rectorados, Ciudad Universitaria presenta una gama de servicios orientada a los estudiantes universitarios como residencias, colegios mayores, polideportivos, o piscinas.

3.1.5. La World Pride 2017 de Madrid

La *World Pride* es el mayor evento del colectivo LGTB del mundo. Es uno de los eventos que más relevancia a nivel internacional ha tenido en redes sociales en los últimos años, por la magnitud de visitantes que la ciudad organizadora recibe de todo el mundo, y por la repercusión que conlleva su propia temática. Este evento tiene como objetivo la promoción a nivel global de aspectos relacionados con el colectivo de lesbianas, gais, bisexuales, y transexuales (LGBT) mediante festivales, conciertos, desfiles y otros tipos de actividades. Redes sociales cómo *Twitter* son vehículos de comunicación bastante utilizados por este colectivo para ganar visibilidad gracias a elementos cómo las tendencias o *trending topics*. La *World Pride* se celebra a nivel mundial desde el año 2000. Su quinta edición se celebró en Madrid en el año 2017, desde el 23 de junio hasta el 2 de julio. El programa más destacado fue la manifestación celebrada el 1 de julio en los Paseos del Prado y de Recoletos. El evento contó además con diversas actividades, cómo conciertos o carreras. El Ayuntamiento de Madrid calculó una asistencia de 2.200.000 personas y un beneficio económico para la ciudad de 115 millones de euros[14].

El barrio de Chueca, ubicado al norte de la Gran Vía, fue el área central de las actividades de la *World Pride*, Este barrio, de ambiente multicultural, es el epicentro de la ciudad para la comunidad LGBT desde los años 1980, década en la que personas pertenecientes a la comunidad LGBT madrileña se hicieron con muchas viviendas y locales, floreciendo las discotecas, pubs, bares, y tiendas orientadas a dicho colectivo. Desde 1986 se celebra anualmente en el barrio de Chueca las fiestas de celebración del Orgullo Gay. En la década de los 90 se desarrolla la manifestación LGBT cómo elemento reivindicativo unido a celebraciones lúdicas[15]. Los festivales se organizan el fin de semana posterior al Día del Orgullo LGBT, el 28 de junio.

[14]https://diario.madrid.es/blog/notas-de-prensa/exito-del-operativo-de-seguridad-movilidad-y-limpieza-del-world-pride-madrid-2017/
[15] http://www.worldpridemadrid2017.com/home/acercade

3.2. Datos utilizados

3.2.1. Datos de Twitter

Los datos de *Twitter* son la base sobre la que se sustenta esta tesis doctoral. La base de datos inicial empleada para para todos los casos de estudio de esta tesis (con la excepción del último, que investiga la percepción del sistema de transporte público) tiene un total de 2.229.253 *tweets*, todos ellos georreferenciados, localizados dentro del Área Metropolitana de Madrid, y producidos por 171.631 usuarios. Estos *tweets* fueron recogidos en un periodo de dos años (desde el 1 de junio de 2016 hasta el 31 de mayo de 2018). Cada *tweet* cuenta con información relativa al identificador de usuario, el nombre de usuario, coordenadas espaciales de latitud y longitud, estampa temporal de fecha y hora, idioma y los *hashtags* que incluye el *tweet*.

La falta de *tweets* con textos relevantes relacionados con el sistema de transporte de Metro provocó la necesidad de una segunda descarga de datos para la realización del quinto caso de estudio de la tesis. Esta segunda base de datos cuenta con un total de 27.603 *tweets* de 12.361 usuarios, recopilados en un periodo de dos meses (desde el 9 de septiembre hasta el 9 de noviembre de 2019). Los *tweets* descargados son respuestas a la cuenta de usuario oficial del Metro de Madrid (*@metro_madrid*), excluyendo los *retweets*. Igual que en la primera base de datos, los *tweets* cuentan un identificador y nombre de usuario, una estampa temporal de fecha y hora, idioma, y otros metadatos. Pero a diferencia de la base de datos anterior, estos *tweets* no están geololocalizados, es decir, no cuentan con datos de coordenadas espaciales.

3.2.2. Datos de uso del suelo

Las localizaciones de los *tweets* fueron cruzadas con datos del uso del suelo, añadiendo información de las actividades primarias de las parcelas desde donde los mensajes fueron enviados. Para ello se ha trabajado con la información del Catastro Nacional de 2017, que contiene datos de alto detalle espacial a nivel de parcelas.

3.2.3. Datos de población residente, empleo y nivel de renta

Los datos de población residente utilizados en esta tesis se han obtenido del Padrón de habitantes del año 2017 del Instituto Nacional de Estadística (INE), tanto a nivel de municipios como de distritos, barrios, y secciones censales. Por su parte, los datos de la localización de puestos de empleo provienen de los registros de la población afiliada a la Seguridad Social del año 2016, del Portal de Datos Abiertos del Ayuntamiento de Madrid (para los distritos de Madrid) y del Instituto de Estadística de la Comunidad de Madrid (para los datos de municipios del Área Metropolitana).

Los datos utilizados para estudios relacionados con el nivel económico del distrito o municipio de residencia han sido obtenidos a partir de los indicadores de renta del año 2016 del Instituto de Estadística de la Comunidad de Madrid (datos de nivel de renta bruta de cada municipio de la Comunidad de Madrid) y los indicadores de renta del año 2015 del *Urban Audit* del Ayuntamiento de Madrid (datos de nivel de renta neta de cada distrito del municipio de Madrid, que fueron transformados en datos de renta bruta).

3.2.4. Encuesta Domiciliaria de Movilidad 2018

Para la verificación de la matriz OD construida en el primer caso de estudio, se emplearon los datos de movilidad obtenidos por el Consorcio de Transportes de Madrid en la Encuesta Domiciliaria de Movilidad (EDM) del año 2018. La información de su fichero de datos incluye códigos de las zonas de origen y destino del área de estudio (a nivel de municipio y distrito), los motivos del viaje, hora de salida, duración del viaje, medio de desplazamiento, título del abono usado y número de viajes. En total, se registraron desplazamientos de más de 85.000 viajeros, de los que se seleccionaron aquellos cuyo motivo de origen o destino está relacionado con los viajes residencia-trabajo. Esta encuesta usa zonificaciones de transporte (unidades socio-morfológicas homogéneas) como referencia espacial para la expansión de datos.

3.2.5. Datos de número de estudiantes universitarios

La página web del Ministerio de Educación, Cultura y Deporte[16] cuenta con un apartado de datos estadísticos del que se han obtenido datos del curso 2017/18, tanto del número de alumnos matriculados por grado, master y doctorado a nivel de universidad y de campus, como datos de número de plazas de oferta, demanda y matricula que oferta cada universidad y campus. Estos datos han sido empleados para validar los datos de *Twitter* y para el cálculo del modelo gravitacional de asignación de estudiantes en el tercer caso de estudio.

3.2.6. Ficheros de transporte privado y público

Para calcular tiempos de viaje desde el municipio de residencia de los usuarios de cada uno de los campus universitarios hasta el destino en transporte privado se han obtenido los tiempos de viaje a partir de la red de carreteras de la compañía *TomTom*. La red es muy detallada, cubre toda el área de estudio y cuenta con una conectividad total.

En el caso del tiempo en transporte público se han usado ficheros *GTFS* suministrados por el Consorcio de Transportes de la Comunidad de Madrid. Estos ficheros cuentan con datos de tiempos viaje del Metro y Metro ligero de Madrid, el servicio de trenes de Cercanías, los autobuses municipales, urbanos, e interurbanos, y permiten el análisis de tiempos de viaje de redes. Además, estos ficheros permiten obtener capas de líneas de los distintos sistemas de transporte público y capas de puntos de las estaciones pertenecientes a cada red.

3.2.7. Datos de uso del Metro de Madrid

Se han descargado datos con el número de usuarios que utilizaron el Metro de Madrid en el año 2018, tanto por estación como por línea de metro, con el objetivo de validar en el quinto caso de estudio la distribución espacial de usuarios de *Twitter* que utilizan el servicio de transporte público. Estos datos fueron descargados desde la página web del Consorcio de Transportes de la Comunidad de Madrid.

[16] https://www.educacionyfp.gob.es/servicios-al-ciudadano/estadisticas/universitaria/estadisticas/alumnado/desde-2015.html

3.2.8. Puntos de interés

Una de las variables exploratorias utilizadas para la elaboración del modelo GWR del quinto caso de estudio es el número de puntos que se hallan a una determinada distancia de una parada de metro. Un punto de interés es un punto de ubicación específica con un valor determinado para la población. Los puntos de interés pueden representar hoteles, centros de salud y hospitales, comercios, centros de educación, museos, monumentos, etc. Para esta tesis se han descargado los puntos de interés del Área Metropolitana de Madrid a partir de datos de 2019 de *OpenStreetMap*.

La tabla 7 recoge las fuentes de datos empleadas a lo largo de esta tesis doctoral.

Tabla 7: Resumen de fuentes de datos utilizadas en la investigación.

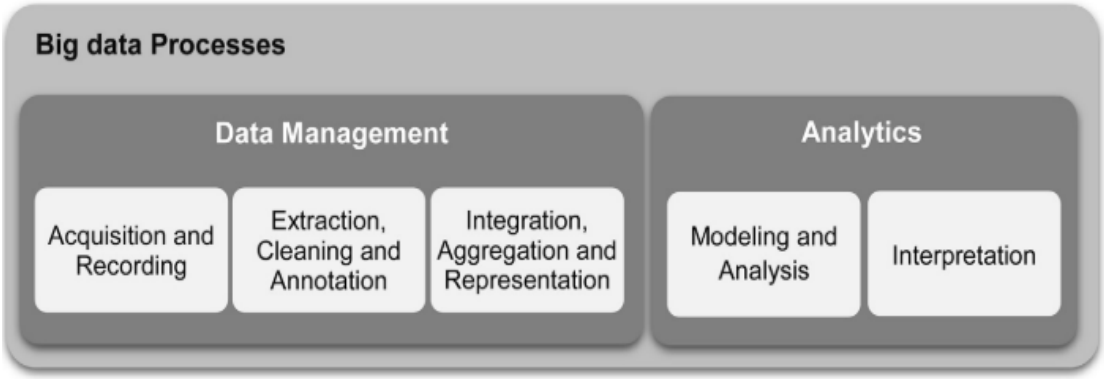
Datos	**Fuente**	**Año**	**Casos de uso donde se usan los datos**
Datos de *Twitter* geolocalizados	API *Twitter*	2016-2018	1, 2, 3, 4
Datos de *Twitter* no geolocalizados	API *Twitter*	2019	5
Datos de usos del suelo	Catastro	2017	1, 2, 3
Datos de población residente (municipio y Comunidad Madrid)	Instituto Nacional de Estadística	2017	1, 3, 5
Datos de población de empleo (municipio Madrid)	Portal de Datos Abiertos Ayuntamiento Madrid	2016	1
Datos de población de empleo (Comunidad Madrid)	Instituto de Estadística Comunidad de Madrid	2016	1
Datos de nivel de renta (municipio Madrid)	*Urban Audit* Ayuntamiento Madrid	2015	3
Datos de nivel de renta (Comunidad Madrid)	Instituto de Estadística Comunidad de Madrid	2016	3
Encuesta Domiciliaria Movilidad	Consorcio Transportes Comunidad de Madrid	2018	1
Tiempos de viaje y redes de transporte público (ficheros *GTFS*)	Consorcio Transportes Comunidad de Madrid	2018	3, 5
Tiempos de viaje transporte privado	*TomTom*	2018	3
Datos número viajeros Metro Madrid	Consorcio Transportes Comunidad de Madrid	2018	5
Datos número estudiantes universitarios	Ministerio de Educación, Cultura y Deporte	2018	3
Puntos de interés	*OpenStreetMaps*	2019	5

Fuente: Elaboración propia.

3.3. Metodología

Trabajar con datos masivos conlleva transformar datos brutos en información con significado añadido. Esta información se vuelve conocimiento al aplicarse en un escenario modelo. El conocimiento generado se emplea para la toma de decisiones en la administración y gestión de la ciudad. Para convertir los datos en conocimiento hay que realizar una serie de pasos que implican la realización de una serie de operaciones de forma secuencial como son la captura de datos, su almacenamiento, depuración, agregación, análisis, visualización, explicación y predicción (Gutiérrez-Puebla et al., 2016). Esta secuencia de pasos será la base a seguir para la metodología de esta tesis doctoral. La Figura 19 muestra un esquema con los diferentes procesos de los que se compone el tratamiento y procesado del *Big Data*.

Figura 19: Pasos de procesamiento del *Big Data*.



Fuente: (Gandomi & Haider, 2015).

En este apartado se relatará la metodología del tratamiento de datos de *Twitter*, es decir, los pasos de almacenamiento, preprocesado y ennriquecimiento de datos. La mayoría de los pasos realizados en el tratamiento del *Big Data* son comunes para los casos de estudio de la tesis. Los últimos pasos del proceso metodológico, el análisis y visualización de los datos, tiene un carácter más específico para cada caso de estudio, por lo que estos pasos se tratarán de forma superficial en este apartado, y se tratarán de forma más profunda en cada uno de los casos de estudio que conforman el siguiente capítulo. Al final del apartado, la Figura 20 muestra un esquema que resume toda la metodología utilizada en esta tesis doctoral.

3.3.1. Descarga y almacenamiento de datos

El ciclo de vida del *Big Data* empieza por la generación de datos, que puede ser de una gran variedad de formas y fuentes. A continuación, estos datos son descargados y recopilados en una infraestructura de almacenamiento para su procesado y análisis (M. Chen et al., 2014). Una de las herramientas más utilizadas para la recolección de datos son las APIs o Interfaces de Programación de Aplicaciones, llaves de acceso a funciones que se pueden utilizar para acceder de manera segura y confiable a los datos de un servicio web provisto por un tercero.

Tal como se comentó en el apartado 3.2.1. de la tesis, se han estado descargando de forma continua datos de *Twitter* durante un periodo de dos años. Esta descarga ha sido posible gracias a la API de *Twitter*. Con la ejecución de un código de *Python* ha sido posible descargar *tweets* con datos geolocalizados de la API que se hallen dentro de un cuadro de coordenadas. Cada esquina del cuadro es un punto con unas coordenadas de latitud y longitud específica con las que se cubrió todo el territorio de la Comunidad de Madrid. Para poder descargar *tweets* de la API se utilizan unos códigos de llaves internas obtenidas con la creación de un perfil de *Twitter Developer*.

El proceso de almacenamiento tiene como objetivo una eficaz administración, seguridad y accesibilidad posterior a los datos. Debido al enorme volumen de datos y a su naturaleza desestructurada, las bases de datos tradicionales se muestran ineficaces para el almacenamiento de datos y el posterior análisis de los mismos. La forma de almacenar los datos del *Big Data* debe adaptase a un posterior manejo de los mismos en formas de trabajar como la computación en nube (donde los datos se almacenan y tratan en una red de internet en lugar de equipos físicos) o la computación paralela (método en el que las bases de datos de gran tamaño se dividen en partes más pequeñas para realizar el mismo cálculo en distintos grupos de datos que posteriormente convergen en un único ordenador).

Para ello, se necesitan bases de datos flexibles que permitan realizar estos procesos de computación con datos desestructurados y variados. Una solución es el uso de bases de datos *NoSQL*, sistemas más flexibles y rápidos que las bases *SQL* (Gutiérrez Puebla, 2018). Este tipo de bases de datos se caracterizan por ser flexibles, facilitar copias simples, ser consistentes, y soportar volúmenes de datos enormes. Un ejemplo es *MongoDB*, una base orientada a documentos, por lo que puede almacenar datos masivos y semiestructurados de *Twitter* en documentos con formato JSON (M. Chen et al., 2014).

El script de *Python* utilizado para la descarga de *tweets* geolocalizados de la API incluye una ruta de salida que almacena y recopila los *tweets* directamente en una carpeta vinculada a una base de datos *MongoDB*. Desde esta base de datos se pueden realizar búsquedas por campos y seleccionar los *tweets* que cumplan con una serie específica de criterios, facilitando su posterior descarga. A continuación, los *tweets* se extrajeron de *MongoDB* mediante otro script con el que se crean archivos en formato JSON. En este script además se seleccionan los campos de los *tweets* que puedan ser útiles para la investigación. De este modo, al extraer los *tweets*, estos cuentan ya solo con los campos de identificación del *tweet* y usuario, campo de fecha, idioma, texto, y coordenadas de latitud y longitud.

A partir de aquí, casi todo el proceso metodológico de esta tesis se ha desarrollado en el software SIG *ArcGIS Pro*. Este SIG de escritorio, de la empresa *ESRI*, es uno de los sistemas más rápidos y potentes para analizar, administrar, visualizar (tanto en 2D como 3D o en vídeos temporales), cartografiar y compartir datos espaciales de forma fácil y productiva. Los *tweets* escritos en los archivos JSON fueron extraídos en capas de puntos mediante una herramienta de conversión a entidades. Las capas de puntos se fusionaron en una sola capa que se almacena en una geodatabase de *ArcGIS*. Ya en este paso se realizó un primer proceso de filtrado de datos: una selección espacial con la que se almacenan solo los *tweets* ubicados dentro del área de estudio.

3.3.2. Limpieza, procesado, y enriquecimiento de datos

Una vez descargados y almacenados los datos, es necesario limpiarlos y filtrarlos para eliminar el mayor ruido y redundancia posible. Para ello se han usado técnicas de extracción, integración, limpieza de información innecesaria o incompleta, optimización, o metadatado. En este paso se detectan datos erróneos, incompletos, redundantes, o imprecisos y se eliminan. Posteriormente se han enriquecido mediante la fusión con multiples fuentes, confiriéndole de nueva información complementaria que no poseían previamente (M. Chen et al., 2014; OECD, 2015).

El paso más delicado del tratamiento de datos de *Twitter* consiste en su limpieza y procesado, ya que conlleva la transformación de datos masivos, desestructurados y con un nivel de ruido importante en una selección datos con una estructura y enriquecimiento de los que se puede extraer información. Una vez incorporados los datos en un SIG, se

procedió a una selección general de filtros para seleccionar usuarios útiles para la investigación y eliminar lo máximo posible el sesgo y ruido de los datos. Los pasos seguidos fueron:

- Primero se eliminaron de la base de datos *tweets* cuyo identificador de usuario perteneciese a cuentas *bot* o robot. Estas cuentas pertenecen a usuarios compulsivos asociados a máquinas que generan mensajes automáticamente. Normalmente los *bot* suelen pertenecer a páginas de noticias o empresas que hacen publicidad de forma masiva y constante. Las cuentas *bot* fueron identificadas como cuentas con más de 1000 *tweets* publicados y con las mismas coordenadas en todos sus *tweets*, o cuentas con más de 10 *tweets* publicados con el mismo contenido semántico en el campo de texto de los *tweets*.
- Una vez eliminadas las cuentas *bot*, se separó la información de la fecha y de la hora del campo de fecha completa de los *tweets*. Después, se extrajeron tanto el valor del número de día de la semana (equivaliendo el lunes a un valor 1, el martes a un valor 2, etc.), como el del número de hora. La mayor parte de los análisis realizados en esta investigación están relacionados con los patrones diarios de la movilidad. Mientras que los usuarios de *Twitter* tienden a tener comportamientos regulares de movilidad durante los días laborables, la movilidad en fines de semana es más errática. Por tanto, se filtraron los *tweets* publicados los viernes por la tarde (en horas posteriores a las 14.00), en fines de semana, y en días festivos (se seleccionaron todos los días festivos del calendario y se les asignó un valor 0 de número de día de la semana). De esta forma, se eliminan potenciales anomalías que puedan afectar al resultado de la investigación, ya que los patrones de movilidad en días de descanso son más aleatorios.
- El siguiente proceso de limpieza consistió en eliminar usuarios cuyos *tweets* tienen siempre localizaciones muy similares. Es decir, se filtraron las cuentas de usuarios sin movilidad espacial, o con una movilidad espacial muy reducida. Para ello, se midieron las distancias entre los *tweets* de cada usuario, y se eliminaron los mensajes de aquellos usuarios con una distancia media menor a 50 metros en la localización de todos sus *tweets.*
- Después, se filtraron las cuentas de usuario con un periodo temporal reducido, eliminando aquellos usuarios que tienen todos sus mensajes concentrados en un

periodo temporal de dos semanas seguidas. De esta forma, se filtran posibles visitantes o turistas, y se intenta asegurar trabajar con población residente.

- Finalmente, se filtraron los usuarios con muy baja actividad en *Twitter*, eliminando aquellos usuarios que hayan publicado menos de 5 *tweets* en total.

Como resultado de todos los procesos de limpieza se obtuvo una base de datos de 1.246.754 *tweets* generados por 73.392 usuarios. Estos filtros se han empleado para casi todas las investigaciones que han empleado datos geolocalizados de *Twitter*, con la excepción del análisis realizado en el apartado 4.4. de la tesis (sobre el impacto de eventos como la *World Pride* 2017), donde solo se efectuó el filtro de cuentas *bot*, y en cambio, se seleccionaron solamente los *tweets* publicados en las semanas seleccionadas para la comparación de distintas casuísticas de movilidad. Igualmente, en el análisis temporal de los datos recogido en el apartado 4.2., se añadió un filtro consistente en eliminar usuarios con un rango horario igual o menor a ocho horas, con el fin de obtener solamente usuarios que hayan estado activos en la mayor parte del día.

A continuación, se amplió la muestra de datos a partir de la descarga de los últimos mensajes publicados por cada usuario con el objetivo de aumentar la precisión espacial y temporal de los movimientos individuales. Esta ampliación permite conseguir los últimos 3.200 *tweets* de cada usuario (tanto geolocalizados como no geolocalizados) con el objetivo de obtener mensajes que no se hubiesen captado en *streaming* (Q. Huang & Wong, 2016). Tras este segundo proceso de descarga, se filtraron los *tweets* no geolocalizados y que no se situasen en el área de estudio o durante el periodo temporal de la muestra original. De esta forma, para el análisis espacial de la movilidad a partir de datos de *Twitter* (capítulo 4.1.), se obtuvo una base de datos final con 1.536.261 *tweets* de 73.392 usuarios (Tabla 8). Se ha escogido este caso de estudio para mostrar una tabla con número de *tweets* y usuarios por cada proceso de limpieza debido a su carácter general, ya que no hay una selección previa de datos ubicados en lugares específicos del área de estudio (capítulos 4.3. y 4.5.) o en un periodo temporal concreto (capítulo 4.4.).

Tabla 8: Proceso de filtrado y expansión de datos de *Twitter*.

Filtro	***Tweets***	***Usuarios***	***Tweets / Usuarios***
Base de datos inicial	2.229.253	171.631	13,0
Cuentas *bot* detectadas	-441.716	-296	1492,3
Tweets publicados en fin de semana o festivo	-396.830	-25.713	1,5
Usuarios con baja movilidad espacial	-62.496	-36.856	1,7
Usuarios con baja movilidad temporal	-35.244	-10.521	3,3
Usuarios con menos de 5 *tweets*	-46.213	-24.853	1,9
Base de datos tras los procesos de filtrado	1.246.754	73.392	17,0
Base de datos final tras la descarga de expansión	**1.536.261**	**73.392**	**20,9**

Fuente: Elaboración propia.

3.3.3. Agregación y enriquecimiento de los datos.

Con relación al desafío que conlleva el uso óptimo de los datos, el enriquecimiento de la información geográfica tiene como objetivo aumentar la calidad y el conocimiento generado a partir de la combinación de distintas fuentes. Destacan el enriquecimiento semántico que otorga mayor número de atributos, y el enriquecimiento geométrico que concede mayor escala y resolución, y mayor facilidad para visualización en 3D (Zipf, 2015). La geolocalización permite enriquecer los datos mediante procesos simples y añadir nuevos atributos a las bases de datos. Estos datos también pueden ser cruzados espacialmente con otras fuentes de datos masivos y con datos espaciales de estadísticas de fuentes tradicionales (Gutiérrez Puebla, 2018).

Anteriormente, para filtrar *tweets* de usuarios publicados en fines de semana o días festivos, se ha comentado que se agregaron los datos de *Twitter* por número de día de la semana y por número de hora del día. Además, los *tweets* también se agruparon por periodos de cuarto de hora. De esta forma, si dos o más *tweets* de un mismo usuario se

hallan en un mismo espacio durante un mismo periodo de cuarto de hora, se actúa como si el usuario hubiese solo publicado un *tweet*. La base de datos fue agregada y enriquecida espacialmente mediante una unión espacial primero con una capa de municipios de la Comunidad de Madrid, y a continuación, con una capa de distritos del municipio de Madrid. De esta forma, se incorporó a cada *tweet* tanto el nombre y geocódigo del municipio (o distrito del municipio de Madrid) desde donde fue publicado. Por último, la base de datos fue enriquecida por el mismo método con datos de usos del suelo. Para ello, se usa una capa de parcelas del Catastro. Esta capa cuenta con un campo que señala el uso del suelo prioritario de cada parcela, por lo que cada *tweet* es vinculado al uso del suelo principal desde donde fue publicado.

3.3.4. Análisis y visualización de los datos.

Con estos datos ya limpios, enriquecidos, organizados y agregados, se puede proceder al análisis de datos. Para ello se emplean métodos estadísticos como análisis de clústeres, correlación, o regresión para analizar datos, concentrar, extraer y refinar datos útiles ocultos, identificar patrones formular análisis, construir modelos y visualizar resultados (M. Chen et al., 2014; OECD, 2015). Las técnicas de estadística espacial son eficaces para tratar el *Big Data* ya que ofrecen capacidades para resumir los datos y expresar medidas de variación e incertidumbre (S. Li et al., 2016).

El último paso del proceso es la visualización e interpretación de los datos. Es en este paso donde el auge del *Big Data* ha revalorizado el uso de los SIG como herramientas que pueden procesar datos espaciales y no espaciales (incluso no estando estructurados) a partir de medios de computación y geovisualización (S. Li et al., 2016). Con los SIG se pueden visualizar e interpretar los datos y presentar los resultados de forma cartográfica, permitiendo generar información valiosa de forma visual con la que generar conocimiento para la toma de decisiones.

En la investigación científica, procede añadir un paso más de validación de los resultados. Este proceso se realiza a partir de la comparación de los resultados obtenidos con los datos de otras fuentes. Es posible realizar la validación de las distribuciones obtenidas de residencia o trabajo y de los procesos de movilidad con datos oficiales como censos o encuestas domiciliarias de movilidad, con datos de tarjetas de transporte, o con grupos de control (Gutiérrez-Puebla et al., 2019).

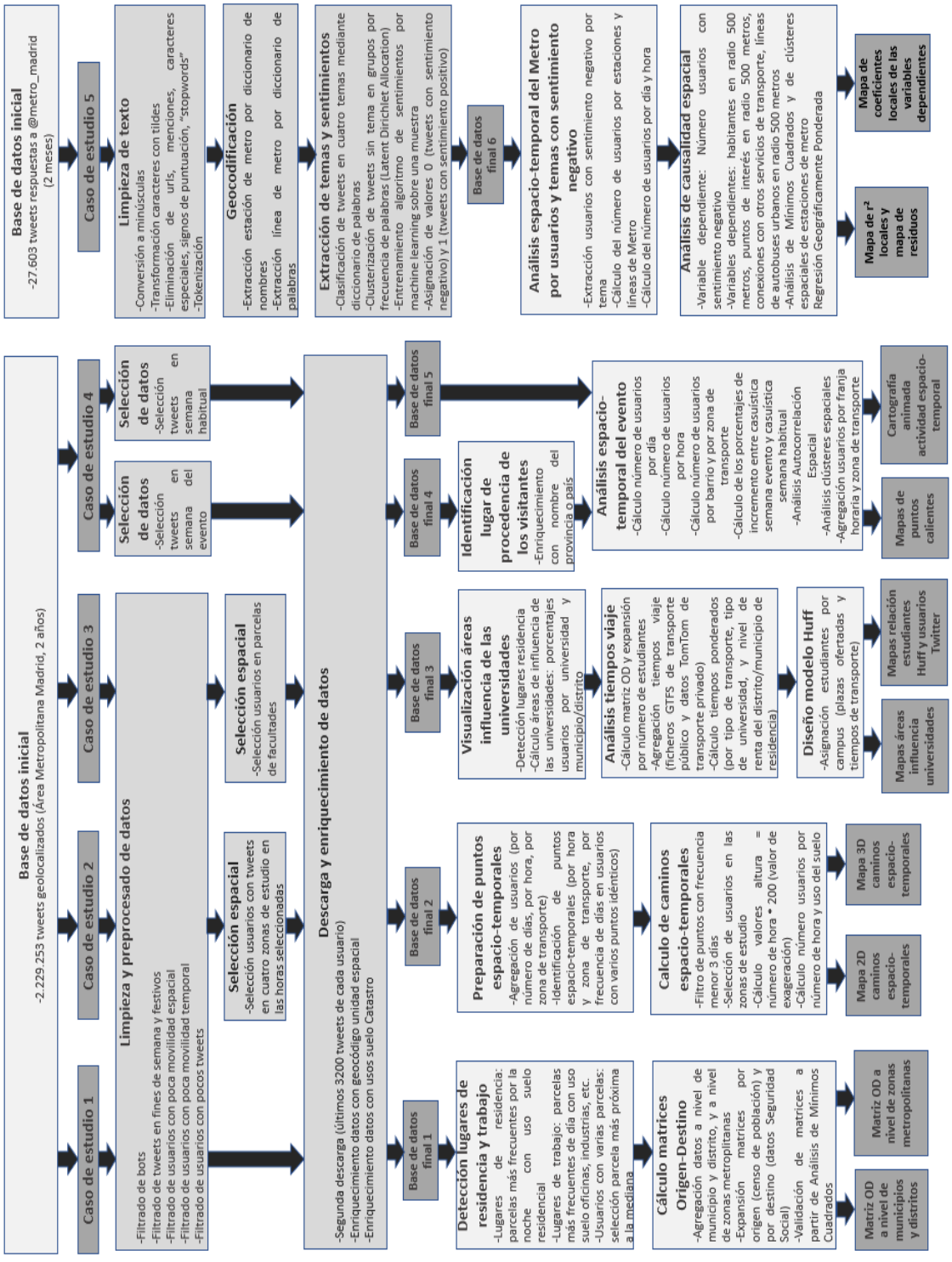
La Tabla 9 resume las técnicas estadísticas y cartográficas empleadas en esta tesis doctoral, y el caso de estudio en el que han sido aplicadas.

Tabla 9: Resumen de técnicas estadísticas y cartográficas utilizadas en la investigación.

Técnica	**Tipo**	**Casos de uso donde se usa la técnica**
Análisis Exploratorio Datos	Geoestadística	1, 2, 3, 4, 5
Matrices OD	Cartografía	1, 3, 4
Técnicas de expansión de datos	Estadística	1, 3
Análisis de Mínimos Cuadrados (OLS)	Geoestadística	1, 3, 5
Mapa de residuos	Cartografía	1, 5
Caminos espacio-temporales	Cartografía	2
Ponderación de tiempos de viaje	Geoestadística	3
Visualización áreas influencia	Cartografía	3
Modelo gravitacional de asignación de usuarios	Geoestadística	3
Análisis semántico	Estadística	4, 5
Análisis de clústeres espaciales (Moran I)	Geoestadística	4, 5
Cartografía de zonas de calor	Cartografía	4
Cartografía interactiva o animada	Cartografía	4
Agrupación de datos por contenido semántico	Estadística	5
Regresión Geográficamente Ponderada (GWR)	Geoestadística	5

Fuente: Elaboración propia.

Figura 20: Esquema metodológico de la tesis doctoral.



Fuente: Elaboración propia.

4. CASOS DE ESTUDIO

4.1. Diseño y validación de matrices de viajes origen-destino a partir de datos de *Twitter*

4.1.1. La movilidad metropolitana a partir de los viajes residencia-trabajo

Un área metropolitana es la suma de una diversidad de actores y actividades que se dan a una mayor escala que en la ciudad, englobando la vida de diferentes ciudades en un ámbito mayor. Se puede interpretar un área metropolitana como un espacio de vida donde la gente se mueve entre diferentes espacios como consecuencia de trabajar en un núcleo central que concentra la oferta de empleo y vivir en una periferia más barata y asequible formada por una serie de ciudades dormitorio.

El desarrollo urbano suele producirse mediante la creación de estructuras centralizadas en las que primero se constituye el área metropolitana como mercado de trabajo, para a continuación consolidarse como mercado de vivienda mediante procesos como la construcción de Programas de Actuación Urbanística (PAU), el establecimiento de servicios en nuevas áreas o la integración de estas zonas en la red de transporte tanto público como privado. Conforme va asentándose, el área metropolitana puede mantener su estructura centralizada o pasar a un proceso de descentralización en el que surgen polos periféricos con mercados propios de trabajo. Por ejemplo, en el Área Metropolitana de Madrid se encuentra una estructura en principio centralizada, donde el municipio de Madrid (especialmente la Almendra Central) genera una gran centralización de puestos de trabajo y recursos, pero a su vez hay una serie de polos periféricos (como el Corredor del Henares, o el cinturón de municipios del sur metropolitano de Madrid) que tienen sus propias características y funcionan como pequeñas áreas metropolitanas (Feria Toribio, 2010; García-Palomares & Gutiérrez-Puebla, 2007; Gutiérrez & García-Palomares, 2007).

Las diferentes fuentes oficiales de los servicios de transporte de la Comunidad de Madrid recogen que el motivo principal de viaje en las áreas metropolitanas es el desplazamiento desde el hogar hasta el trabajo. Los flujos de movilidad residencia-trabajo son esenciales para entender la estructura y funcionamiento básicos de un área metropolitana. La propia movilidad residencial es un factor del que parten el desarrollo físico y la organización metropolitanas, obviando los límites administrativos municipales que se quedan obsoletos al configurar el área metropolitana una unidad mayor (Feria Toribio, 2010). A la vez, los

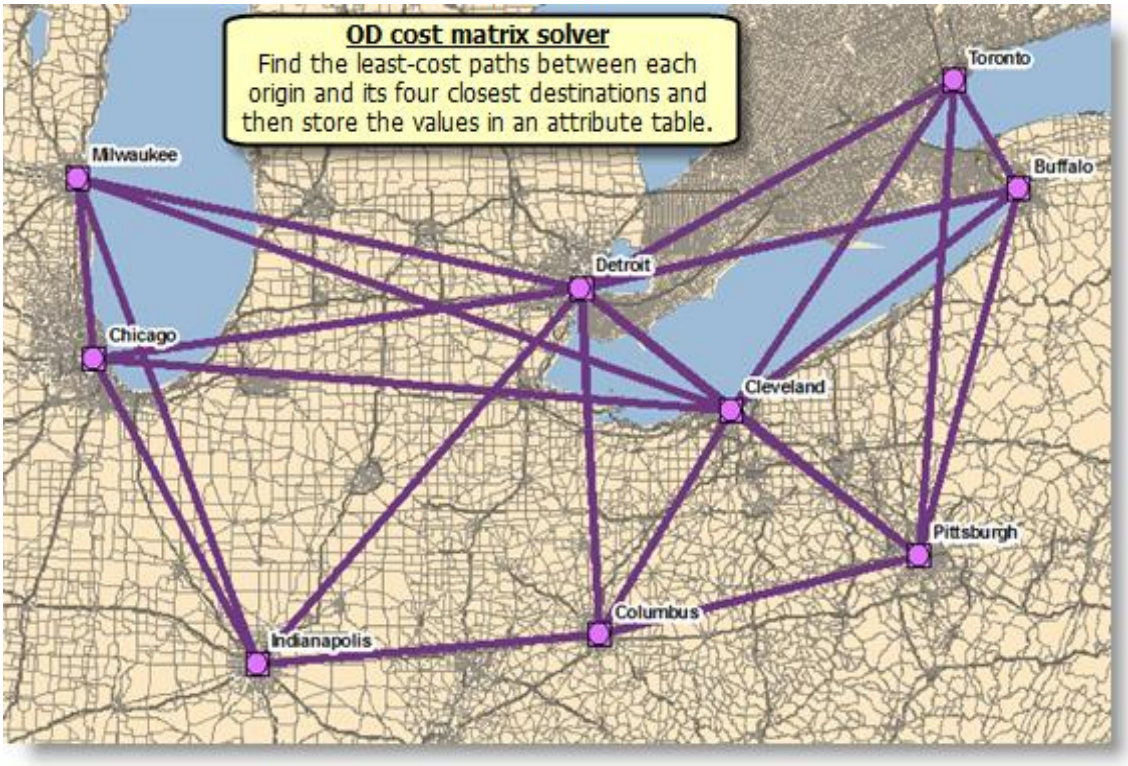
flujos de movilidad están condicionados por diferentes factores como la ubicación de infraestructuras y servicios, la conexión con los servicios de transporte, o los usos del suelo de las diferentes zonas. Comprender la movilidad residencia-trabajo cuenta con un gran valor económico y práctico, ya que tanto los espacios de residencia como de trabajo son lugares clave para el diseño y el planteamiento urbano. Poder comprender como los trabajadores se mueven durante el día es importante para diseñar y administrar los sistemas de transporte y espacios públicos (Pajević & Shearmur, 2017).

Las matrices de flujos Origen-Destino (OD) de viajes son una herramienta indispensable para la evaluación de los planes y proyectos de infraestructuras y servicios de transporte (Gutiérrez-Puebla et al., 2019). Esta información contribuye al modelado del transporte y la optimización del uso de redes, la realización de diagnósticos de movilidad, la previsión de rutas, el planeamiento de servicios, y la predicción de demanda de viajes (Gao et al., 2014). En la investigación de la movilidad residencia-trabajo, las matrices OD calculan además de los viajes, los tiempos y distancias de los mismos a partir de una serie de puntos de origen correspondientes con la demanda potencial de la distribución de la población en los puntos de gravedad de una serie de entidades territoriales, y una serie de puntos de destino equivalentes a la oferta potencial de trabajo que atrae a la población (Rashidi et al., 2017). Los principales indicadores de movilidad que pueden obtenerse a partir de las matrices OD son el número de viajes con origen en cada zona y sus destinos, los viajes con destino en cada zona y sus orígenes, el número de viajes por persona, las distancias y tiempos medios de viaje, los viajes según motivos o actividades, y la distribución temporal de los viajes realizados (Gutiérrez-Puebla et al., 2019).

Tradicionalmente se han empleado matrices OD como parte del modelo clásico de cuatro pasos para la modelización del transporte. El primer paso de este modelo consiste en la generación de viajes, es decir, en obtener la cantidad de viajes generados y atraídos por cada una de las zonas. El segundo paso, la distribución de viajes, es cuando las matrices OD se diseñan para visualizar espacialmente el número de viajes entre zonas. A partir de las matrices OD se puede calcular el tercer paso de selección del modo de transporte (construcción de matrices OD según el tipo de transporte utilizado), y el cuarto paso de selección de ruta (asignación de la cantidad de viajes que pasan por cada una de las parejas origen-destino en los diferentes modos de transporte con el objetivo de determinar que ruta tomarán los viajeros).

Visualmente, las matrices OD suelen representarse en capas de líneas rectas que conectan una serie de puntos de origen con varios puntos de destino. El software *ArcGIS* permite el diseño de matrices OD, en este caso de tiempos de viaje, que encuentran y miden las trayectorias de menor coste a lo largo de una red predefinida. Estas líneas no siguen la red, sino que almacenan información acerca de los tiempos o distancias de viaje en la base de datos adjunta[17]. La Figura 21 muestra un ejemplo de una visualización sencilla en *ArcGIS*. Al poder tratarse como un *shapefile*, es posible incorporar datos externos en una matriz OD como el número de viajeros que realizan un trayecto señalado por una línea.

Figura 21: Ejemplo de Matriz OD de tiempos de viaje.



Fuente: ArcGIS.

Tradicionalmente se han empleado encuestas domiciliarias o conteos de tráfico para obtener matrices de viajes. Sin embargo, este tipo de fuentes son caras, lentas de obtener, estáticas, y según los casos pueden usar pequeñas muestras de población (Iqbal, Choudhury, Wang, & González, 2014). Ante este paradigma, los datos de las redes sociales en general, y de *Twitter* en particular, se antojan como nuevas fuentes para el cálculo de matrices de viajes OD. Las nuevas fuentes de datos basadas en las TIC

[17] https://desktop.arcgis.com/es/arcmap/latest/extensions/network-analyst/od-cost-matrix.htm

permiten trabajar con datos de alta resolución espacio-temporal y acceder a muestras mayores de población, en algunos casos de forma gratuita. A partir de estos datos es posible conocer los lugares en los que el usuario realiza sus actividades, analizar sus patrones de movilidad, y generar matrices OD tanto totales como por franjas horarias (Gutiérrez-Puebla et al., 2019).

La mayor parte de los trabajos basados en TIC para obtener matrices OD de viajes han usado datos CDR de telefonía móvil (C. Chen et al., 2016). Sin embargo, las compañías de telefonía móvil son muy reacias a ceder sus datos. Además, el uso de estos datos conlleva una menor resolución espacial al tener que trabajar con polígonos de *Voronoi*. Teniendo en cuenta estas dos limitaciones principales, en este caso de estudio se busca validar *Twitter* como alternativa a los datos de telefonía en el diseño de matrices OD de viajes, aprovechando que los datos de *Twitter* son gratuitos y que podemos trabajar con la localización exacta de los *tweets* y asociar la información a zonas de transporte, lo cual es de gran interés para los gestores de movilidad.

En este caso de estudio se propone como objetivo investigar los patrones de movilidad residencia-trabajo del Área Metropolitana de Madrid a partir de los datos descargados de *Twitter*, y visualizar estos flujos mediante matrices OD. En el tratamiento de los datos de *Twitter* para la obtención de las matrices de viajes se han realizado algunas mejoras metodológicas como el uso de datos de usos del suelo del Catastro. Estos datos cuentan con alto detalle espacial, ofreciendo mayor precisión a la hora de detectar los lugares de residencia y trabajo. Además, se han evaluado distintos factores de expansión sobre la muestra. Por un lado, se ha trabajado con expansiones a partir de los orígenes de los viajes, utilizando datos de población según el lugar de residencia de fuentes oficiales (padrón de habitantes). Por otro lado, se han expandido las matrices a partir de los datos de destino de viajes, utilizando datos del lugar de trabajo (a partir de los registros de la Seguridad Social). En los análisis se han utilizado dos agregaciones espaciales, la primera con un mayor detalle espacial, donde se trabaja a nivel de municipios y distritos en la ciudad central de Madrid, y la segunda con un detalle espacial menor, utilizando una división en zonas metropolitanas. Los resultados han sido validados a partir de la comparación con los datos de la Encuesta Domiciliaria de Movilidad (EDM) realizada por el Consorcio de Transportes de Madrid en el año 2018, lo que permite evaluar la calidad de las matrices obtenidas mediante datos de *Twitter* y la sensibilidad de los mismos tanto al tipo de factor de expansión a utilizar como al nivel de desagregación espacial.

4.1.2. Metodología específica para el diseño de matrices OD a partir de datos de Twitter

4.1.2.1. Identificación del lugar de residencia y trabajo de los usuarios

Tras limpiar y enriquecer de información espacial y temporal a cada *tweet* de la base de datos, es posible identificar el lugar de residencia y trabajo o estudio de cada uno de los usuarios. Para ello, los *tweets* se diferenciaron entre diurnos (mensajes envíados entre las 08:30 y las 20:30) y nocturnos (los *tweets* restantes). Los *tweets* enviados por la noche fueron asociados inicialmente al lugar de residencia del usuario, y los *tweets* publicados durante el día al lugar de trabajo. Esta metodología ha sido tradicionalmente usada en trabajos de movilidad usando datos de telefonía móvil, mediante los CDR (Ahas et al., 2010; Alexander et al., 2015; Picornell et al., 2015).

Para identificar el lugar de residencia de los usuarios, se ha trabajado con las parcelas desde donde cada usuario ha enviado *tweets* con mayor frecuencia por la noche. Al cruzar los *tweets* con datos de usos del suelo, se pueden considerar solamente las parcelas cuyo uso es residencial (Figura 22). De este modo, se han eliminado posibles residencias erróneamente asociadas a trabajadores nocturnos o lugares de ocio. Cuando se han encontrado usuarios con más de dos parcelas residenciales con *tweets* por la noche, y donde la mayor frecuencia tiene el mismo número de *tweets*, se ha usado la herramienta de centro de la mediana incluida en *ArcGIS*, y de esas parcelas se seleccionó la más próxima al centro de mediana calculado. En los casos de usuarios con solo dos parcelas y el mismo número de *tweets* en cada una de ellas, se descartan ya que no se puede decidir cuál de las parcelas podría ser su residencia.

El mismo procedimiento fue seguido para determinar la localización de los lugares de trabajo de los usuarios, esta vez trabajando con *tweets* publicados durante el día. En este caso, se seleccionaron las parcelas con *tweets* cuyo uso principal de suelo está relacionado con el trabajo (oficinas, industrias, etc.), eliminando de este modo *tweets* diurnos realizados desde espacios residenciales y en potenciales espacios de ocio. Finalmente, las parcelas donde el usuario ha publicado *tweets* de forma más frecuente fueron definidas como el lugar de trabajo. Una vez más, el centro de mediana se empleó para establecer una parcela individual en el caso de que un usuario tuviese más de dos parcelas con el mismo número de *tweets*.

Figura 22 : Ejemplo de detección de parcela de residencia por moda.

Fuente: Elaboración propia.

Esta metodología sigue en parte el procedimiento de (Ahas et al., 2010) de detectar puntos de anclaje de residencia y trabajo con datos CDR. Esto significa que el método utilizado aquí está basado en la metodología empleada habitualmente con datos de telefonía móvil, pero en este caso se usan datos de uso del suelo para mejorar la definición de la detección de residencias y lugares de trabajo. Mientras que (Ahas et al., 2010) utilizaron células de redes para señalar la localización de usuarios de teléfono móvil, los datos del uso de suelo son una excelente alternativa para enriquecer datos de *Twitter* de forma precisa, mostrando si un usuario ha enviado un *tweet* desde su residencia, lugar de trabajo, o centros de ocio como tiendas.

Por último, los lugares de residencia que se han obtenido a nivel de parcela fueron agregados a nivel de municipio o distrito en caso de la ciudad de Madrid para posteriores análisis, mitigando así los problemas de privacidad que pudiese ocasionar la detección de parcelas de residencia.

4.1.2.2. Construcción y expansión de las matrices OD

Una vez definidas las localizaciones del lugar de residencia y del trabajo o estudio, se utilizaron los identificadores de usuario para calcular las matrices de viajes según distritos y municipios a partir de las relaciones entre el lugar de residencia (origen) y el lugar de trabajo (destino). El resultado es una matriz con un total de 20.744 usuarios. Las matrices se obtienen para 21 distritos dentro del municipio de Madrid y 49 municipios del área metropolitana. El total de relaciones en la matriz es de 4.900 relaciones (70 orígenes por 70 destinos).

La siguiente fase fue la expansión de la matriz de viajes a partir de la muestra de usuarios de *Twitter*, a una matriz que represente la población entera en edad de trabajo (19-55 años) en el área Metropolitana de Madrid. El objetivo de este paso es expandir los datos obtenidos de la muestra con el fin de obtener estadísticas de movilidad que representen al total de la población en el área de estudio (Gutiérrez-Puebla et al., 2019). Dos procesos de expansión fueron utilizados para este fin:

El primer método, a partir de datos de origen de viaje, se basó en datos de población del Censo oficial, teniendo en cuenta la población residente en un rango de edad entre 19-55 años de acuerdo con los datos del censo oficial, empleando la siguiente fórmula:

$$T_{ij}^{e} = T_{ij}.\frac{p_i}{\tilde{p}_i}, \forall i \in N,$$

donde $\frac{p_i}{\tilde{p}_i}$ son los pesos calculados para cada distrito y municipio N, basándose en el radio entre el número total de habitantes de datos del Censo p_i, y las muestras de usuarios de *Twitter* $\tilde{p}_i$ obtenidos en la matriz. Estos pesos fueron multiplicados por el valor de los viajes de *Twitter* obtenidos en cada flujo T_{ij}, siendo el resultado los flujos expandidos T_{ij}^{e}.

El segundo factor de expansión fue calculado empleando datos relacionados con los lugares de trabajo o destinos. Este número fue dividido por el número de trabajadores registrados en el Instituto Nacional de la Seguridad Social en cada distrito o municipio, empleando la siguiente fórmula:

$$T_{ij}^{e} = T_{ij}.\frac{w_j}{\tilde{w}_j}, \forall j \in N,$$

donde $\frac{w_j}{\tilde{w}_j}$ son los pesos calculados para cada municipio y distrito N, basándose en el radio entre el número total de trabajadores (registrados en la Seguridad Social) w_j en el

municipio o distrito *j*, y el número de trabajadores de *Twitter* $\widetilde{w}_j$ obtenidos en el municipio o distrito *j*.

En definitiva, se obtuvieron dos matrices expandidas, la primera a partir del origen y usando los datos de residentes según el censo de población, y la segunda a partir del destino y usando como factor de expansión los trabajadores identificados por la Seguridad Social. A partir de las matrices de relaciones entre distritos y municipios se agregaron los datos para obtener matrices de viajes a nivel de grandes zonas metropolitanas.

Para verificar los resultados obtenidos se han calculado los coeficientes de determinación (correlación de Pearson al cuadrado o r^2) entre los datos de residencia hallados de *Twitter* y los datos de población del censo, y los lugares de trabajo detectados en *Twitter* y los datos del registro de Seguridad Social. A continuación, las matrices obtenidas tanto a nivel de distrito y municipio como a nivel agregado con las distribuciones de viajes de la EDM del Consorcio de Transportes. Los coeficientes r^2 se han calculado utilizando los viajes de la EDM cuyos motivos eran trabajo mediante la siguiente fórmula:

$$T_{ij}^{EMD} = \alpha + \beta . T_{ij}^{e} + \varepsilon$$

donde T_{ij}^{EMD} son los flujos de la EDM y T_{ij}^{e} son los flujos expandidos calculados mediante *Twitter*. Por último, los residuos ε fueron calculados mediante la regresión lineal de Mínimos Cuadrados Ordinarios (OLS), y cartografiados para visualizar las relaciones que presentan las mayores desviaciones sobre los datos de la EDM.

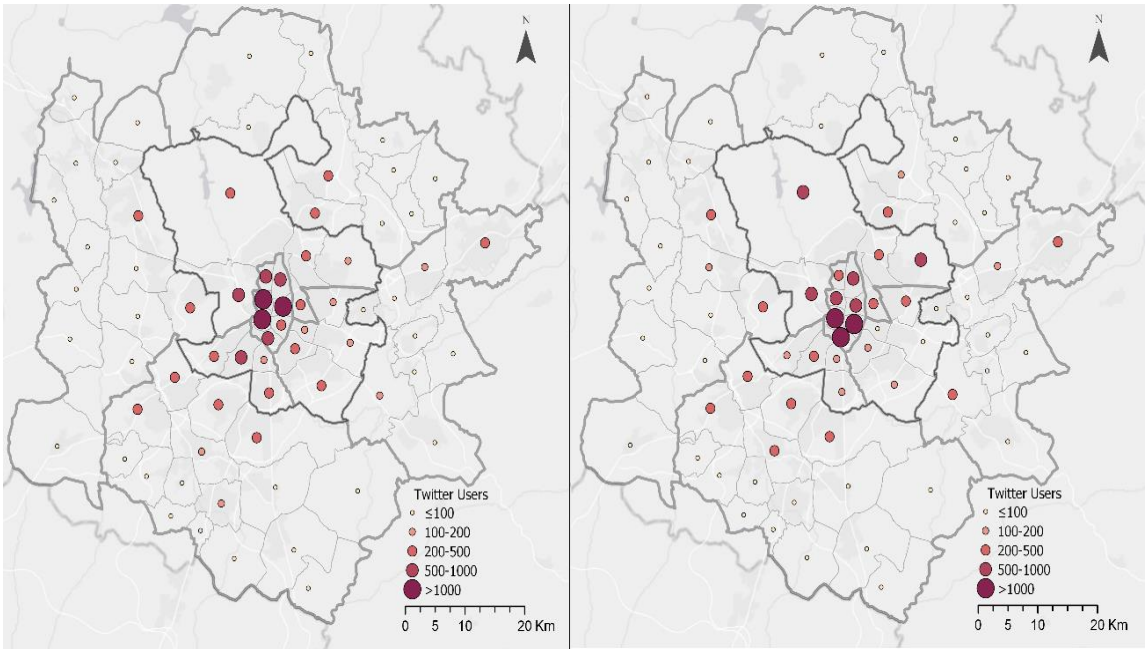
4.1.3. Distribución de los usuarios de Twitter en el espacio según el lugar de residencia y trabajo o estudio

Como se ha señalado, en total se han obtenido matrices OD de movilidad de 20.744 usuarios de *Twitter*, para los que se ha identificado tanto el municipio o distrito de residencia como de trabajo o estudio. Estos usuarios representan casi un 0,7% de la población del área metropolitana. Respecto al lugar de residencia, la mayor cantidad de residentes encontrados a partir de datos de *Twitter* se localizaron en el municipio de Madrid, en especial en los distritos centrales de la Almendra Central (los distritos Centro, Chamberí y Salamanca, zonas dinámicas, con una importante población joven), y en menor grado en los distritos del sur de la ciudad como Carabanchel o Latina (áreas residenciales que conforman los barrios más poblados de la ciudad). Fuera del municipio

de Madrid, los municipios del norte (Alcobendas, San Sebastián de los Reyes) y sur del área metropolitana adyacentes a Madrid (Getafe, Leganés, Alcorcón) destacan como ciudades dormitorio. Los municipios alejados de la capital tienden a tener población pequeña y envejecida (Figura 28 y Tabla 10).

En cuanto al número de lugares de trabajo o estudio identificados, nuevamente los distritos centrales del municipio de Madrid (especialmente Centro, Retiro, Arganzuela, y Chamartín), al ser distritos comerciales, turísticos, y de oficinas destacados, están mejor representados. También destacan los distritos del norte de la ciudad, especialmente los distritos de Barajas (ubicación del aeropuerto de Madrid) y Fuencarral-El Pardo (zona de oficinas junto al distrito de Chamartín). Fuera de la capital, los municipios con mayor cantidad de trabajadores son los municipios del sur (Getafe, Leganés, Alcorcón, Fuenlabrada, Móstoles) y el este del área metropolitana (especialmente Alcalá de Henares y Rivas Vaciamadrid), orientados a la industria, y que conforman dos polos periféricos de atracción de usuarios. De nuevo, los municipios más alejados tienden a tener pocos usuarios (Figura 23 y Tabla 11).

Figura 23: Número de usuarios de *Twitter* por lugar de residencia (izquierda) y de trabajo (derecha).



Fuente: Elaboración propia a partir de datos de *Twitter*.

Tabla 10: Distribución de la población residencial según zonas metropolitanas.

Zonas	**Residentes (*Twitter*)**	**Residentes (Padrón)**	**Porcentaje**
Almendra Central	12.242	503.050	2,43
Distritos norte	1.678	382.191	0,44
Distritos Suroeste	1.188	381.012	0,31
Distritos Sureste	909	331.670	0,27
Sur metropolitano	2.137	670.412	0,32
Este metropolitano	853	330.085	0,26
Norte metropolitano	700	196.703	0,36
Oeste metropolitano	1.037	272.134	0,38
Total	20.744	3.067.257	0,68

Fuente: Elaboración propia a partir de datos de *Twitter*.

Tabla 11: Distribución de los lugares de trabajo según zonas metropolitanas.

Zonas	**Trabajadores (*Twitter*)**	**Trabajadores (Seguridad social)**	**Porcentaje**
Almendra Central	11.415	995.418	1,15
Distritos norte	3.020	486.648	0,62
Distritos Suroeste	769	184.728	0,42
Distritos Sureste	760	247.142	0,31
Sur metropolitano	2.077	348.601	0,60
Este metropolitano	1.043	202.823	0,51
Norte metropolitano	624	225.288	0,28
Oeste metropolitano	1.036	263.771	0,39
Total	20.744	2.954.419	0,70

Fuente: Elaboración propia a partir de datos de *Twitter*.

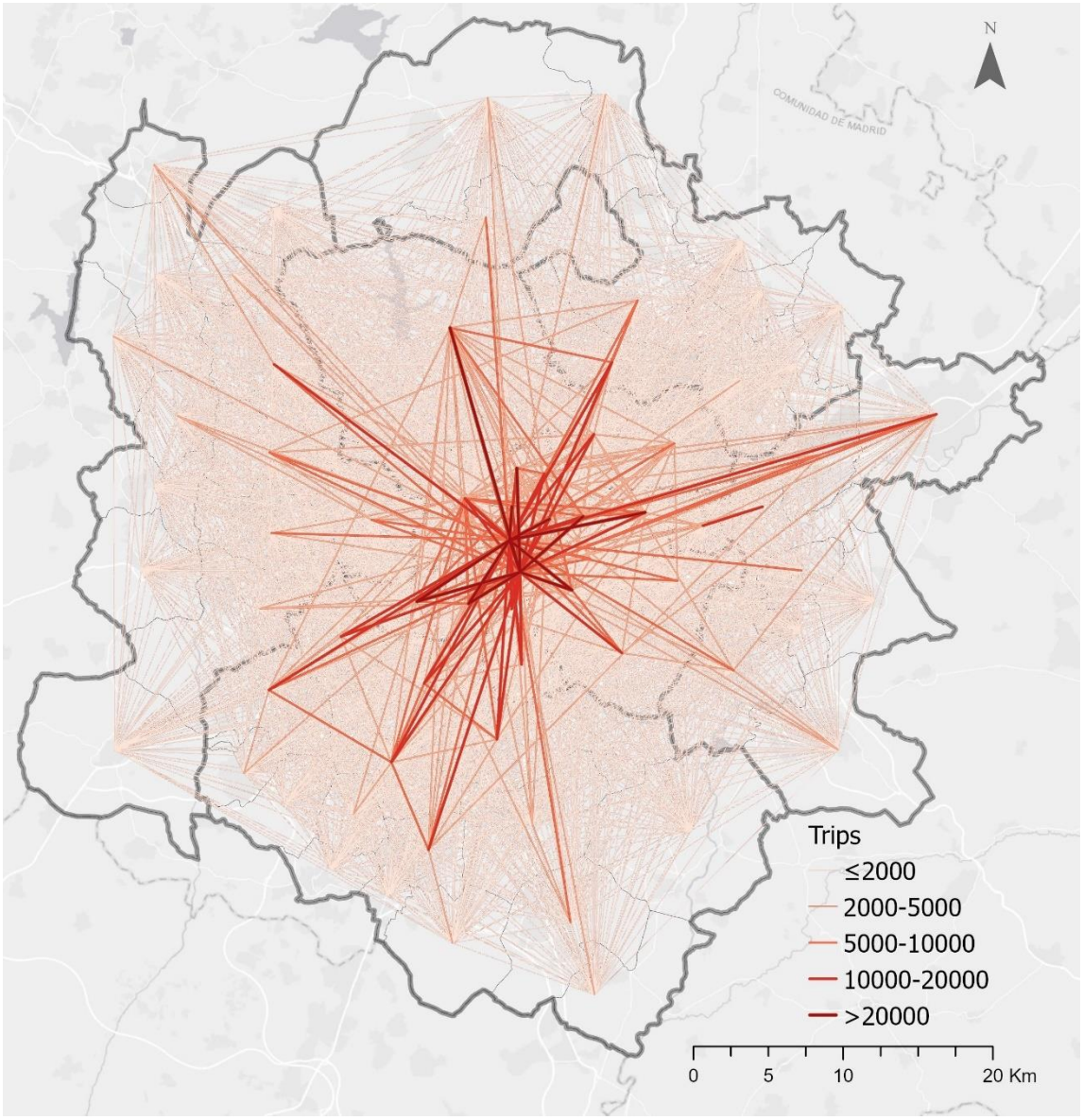
4.1.4. Visualización de matrices OD a partir de datos de Twitter

La Figura 24 representa la maraña de flujos de viajes de la matriz origen-destino entre municipios y distritos del área metropolitana, a partir de la agregación por datos de origen basados en la población del Censo. Destacan los flujos centrípetos, con origen en los distritos periféricos (principalmente los distritos del sur de Madrid) y los municipios metropolitanos (especialmente del Corredor del Henares ubicado al este del área metropolitana, y del cinturón de municipios del sur metropolitano), y destino en la Almendra Central. Entre los viajes periferia-periferia destacan algunas conexiones entre municipios del área metropolitana, por ejemplo, entre los grandes municipios industriales y logísticos del sur metropolitano (el cinturón formado por los cinco grandes municipios del sur) y del este (Corredor del Henares), o las atracciones de nuevas zonas de actividad terciaria en el Norte (municipios de Alcobendas y San Sebastián de los Reyes) y el oeste del área metropolitana. Obviamente la intensidad de las relaciones disminuye en los municipios más periféricos del área metropolitana.

Otra matriz fue expandida mediante datos de trabajadores de la Seguridad Social en áreas de destino. En esta matriz, predomina la concentración de viajes con destino al centro de la ciudad, y los flujos entre el centro y los distritos del este del municipio de Madrid. Cómo se explicará luego, la correlación con la matriz de datos de transporte es baja, por lo que esta matriz fue descartada por su poca precisión.

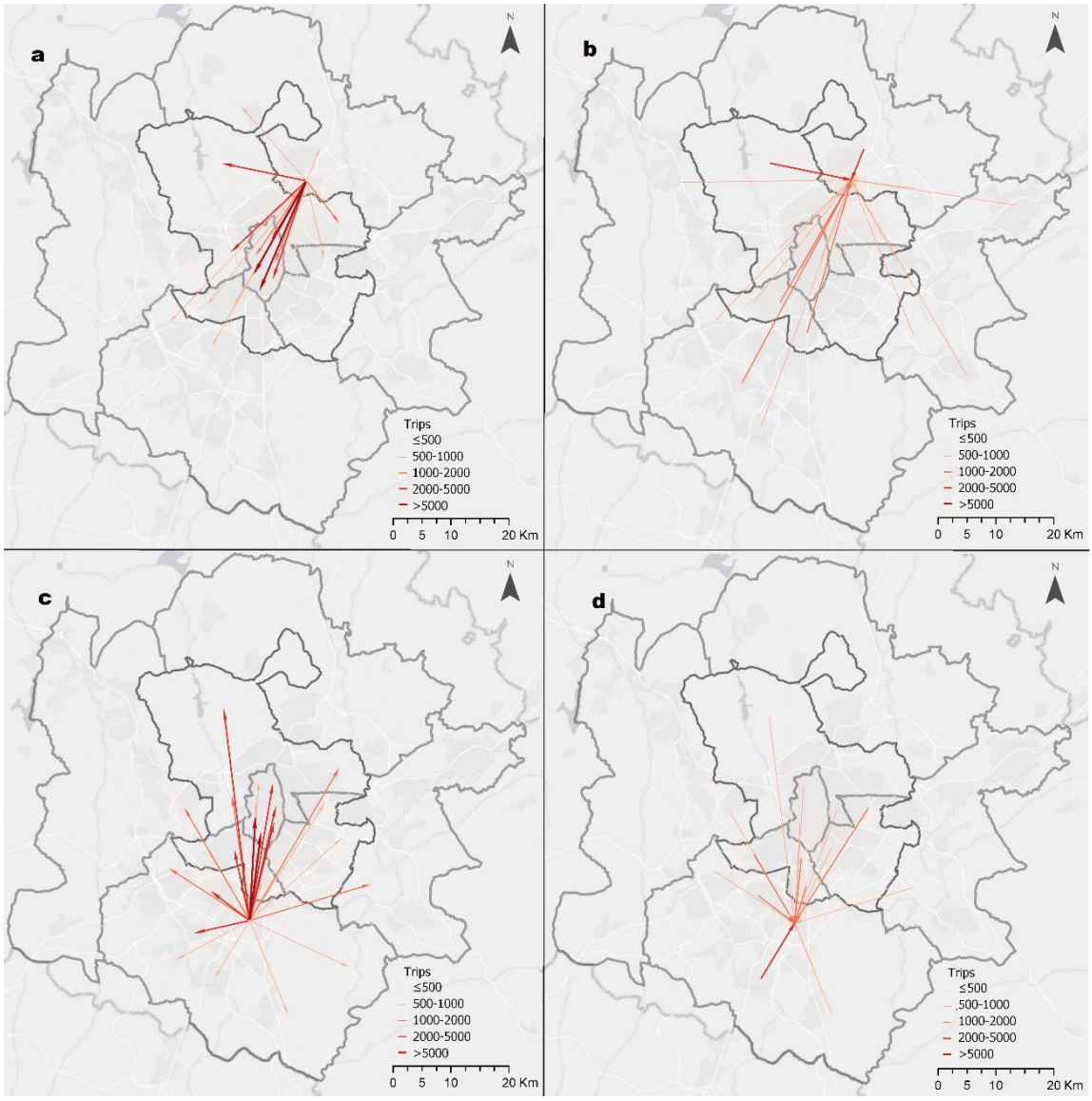
De esta maraña de relaciones, puede obtenerse información de las generaciones y atracciones de cada uno de los distritos o municipios. La Figura 25 muestra como ejemplo dos municipios ubicados en el sur y el norte del área metropolitana. Se trata de dos espacios de marcado carácter residencial, pero que en los últimos años han recibido actividades y empresas como consecuencia de los procesos de descentralización del municipio de Madrid. Tanto las generaciones como sobre todo las atracciones del municipio de Alcobendas, en el norte metropolitano, un espacio terciario y con nivel alto de renta, se relacionan con los espacios de su entorno y con el municipio de Madrid y zonas del sur y el este metropolitano. Mientras, los flujos del municipio de Getafe, espacio principalmente residencial del sur metropolitano poblado por trabajadores con nivel medio de renta, muestran su condición como ciudad dormitorio y tienen una relación mayor con los espacios de su entorno y otros municipios del sur metropolitano.

Figura 24: Matriz de flujos de viajes a partir de datos de *Twitter* a nivel de municipios y distritos del Área Metropolitana de Madrid.



Fuente: Elaboración propia a partir de datos de *Twitter*.

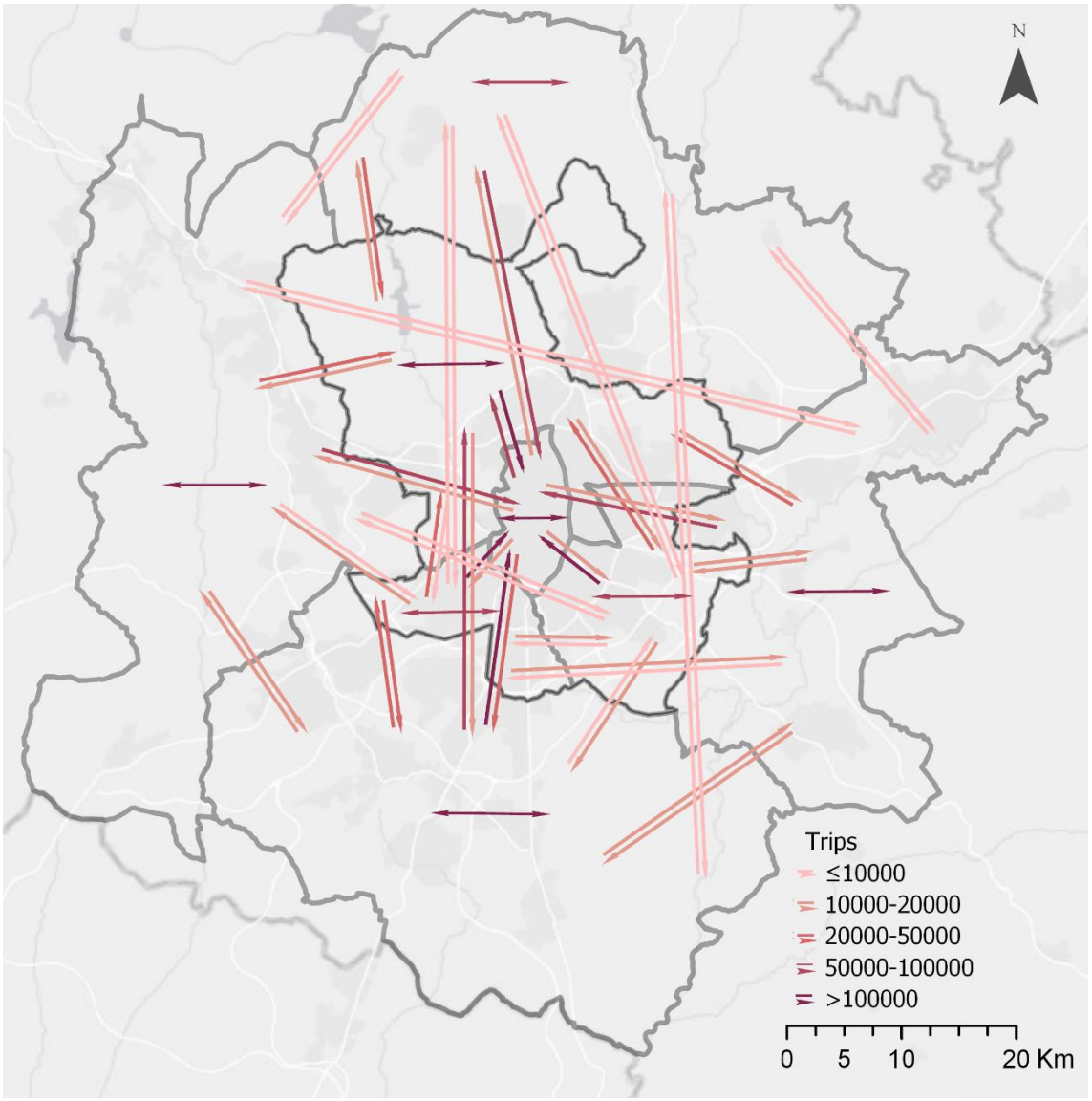
Figura 25: Generaciones y atracciones de Alcobendas y Getafe (a: generaciones de Alcobendas, b: atracciones de Alcobendas, c: generaciones de Getafe, d: atracciones de Getafe).



Fuente: Elaboración propia a partir de datos de *Twitter*.

La matriz, al ser agregada y simplificada para grandes zonas metropolitanas, ayuda a ver de un modo más sencillo las relaciones entre las distintas zonas del área de estudio. Entre los flujos por zonas metropolitanas destacan, por un lado, el volumen de los viajes intrazonales, mostrando las fuertes relaciones entre municipios y distritos próximos, y por otro las atracciones de la Almendra Central. Los espacios centrales tienen una gran importancia en el Área Metropolitana, al contar con un importante volumen de viajes que tiene como destino el centro de la ciudad (Figura 26 y Tablas 12 y 13).

Figura 26: Matriz de flujos de viajes a partir de datos de *Twitter* a nivel de zonas metropolitanas de Madrid.



Fuente: Elaboración propia a partir de datos de *Twitter*.

Tabla 12: Número de flujos totales de viajes a partir de datos de *Twitter* en las zonas metropolitanas de Madrid.

Totales									
	Almendra Central	Distritos Norte	Distritos Sudoeste	Distritos Sudeste	Municipios Sur	Municipios Este	Municipios Norte	Municipios Oeste	Total
Almendra Central	339149	72842	15094	13354	22596	15624	10097	15678	504434
Distritos Norte	184937	126909	5513	13083	12649	11981	14465	12492	382029
Distritos Sudoeste	172575	39938	85021	12286	36571	11778	8372	14519	381060
Distritos Sudeste	138181	43237	7702	99248	13048	15177	6520	8455	331568
Municipios Sur	169942	62502	20908	9724	364165	14363	7436	19685	668725
Municipios Este	74446	35679	4850	10823	14405	174004	8460	7422	330089
Municipios Norte	71320	27988	4677	3920	8883	8130	66454	5315	196687
Municipios Oeste	76491	43342	4936	5923	12654	5544	4382	118785	272057
Totales	1227041	452437	148701	168361	484971	256601	126186	202351	3066649

Fuente: Elaboración propia a partir de datos de *Twitter*.

Tabla 13: Número de flujos porcentuales de viajes a partir de datos de *Twitter* en las zonas metropolitanas de Madrid.

Porcentajes									
	Almendra Central	Distritos Norte	Distritos Sudoeste	Distritos Sudeste	Municipios Sur	Municipios Este	Municipios Norte	Municipios Oeste	Total
Almendra Central	67,23	14,44	2,99	2,65	4,48	3,10	2,00	3,11	100,00
Distritos Norte	48,41	33,22	1,44	3,42	3,31	3,14	3,79	3,27	100,00
Distritos Sudoeste	45,29	10,48	22,31	3,22	9,60	3,09	2,20	3,81	100,00
Distritos Sudeste	41,68	13,04	2,32	29,93	3,94	4,58	1,97	2,55	100,00
Municipios Sur	25,41	9,35	3,13	1,45	54,46	2,15	1,11	2,94	100,00
Municipios Este	22,55	10,81	1,47	3,28	4,36	52,71	2,56	2,25	100,00
Municipios Norte	36,26	14,23	2,38	1,99	4,52	4,13	33,79	2,70	100,00
Municipios Oeste	28,12	15,93	1,81	2,18	4,65	2,04	1,61	43,66	100,00
Total	40,01	14,75	4,85	5,49	15,81	8,37	4,11	6,60	100,00

Fuente: Elaboración propia a partir de datos de *Twitter*.

4.1.5. Calidad y validación de los datos obtenidos

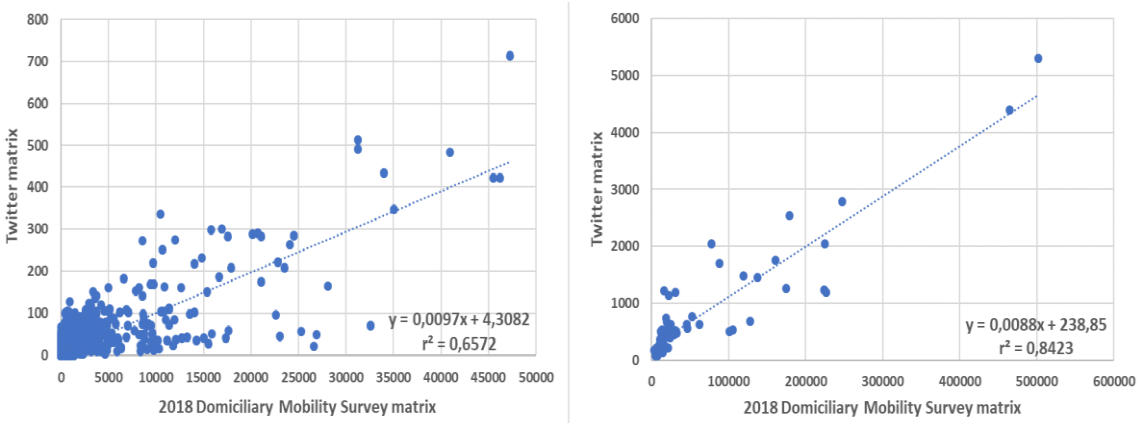
Se puede valorar la calidad de la muestra obtenida en *Twitter* a partir del coeficiente de determinación. Este coeficiente se obtiene entre el número de usuarios residentes detectados en *Twitter* y los datos oficiales de población. El valor r^2 obtenido es relativamente alto con un 0,72. Del mismo modo, el valor r^2 resultado de correlacionar los lugares de trabajo de *Twitter* y los datos de trabajadores de la Seguridad Social marca un 0,70. En este sentido, los valores de la muestra están relativamente bien relacionados con la distribución de residentes y trabajadores del área de estudio, excepto en el caso de

los distritos centrales del municipio de Madrid, donde la muestra de usuarios es mucho mayor de lo esperado, teniendo en cuenta su población residente. Este desajuste en los espacios centrales se debe a que *Twitter* incluye grupos que no son tenidos en cuenta en estadísticas oficiales, cómo extranjeros no registrados y visitantes o turistas no recogidos en los datos oficiales. Además, los ciudadanos de Madrid cuentan con una alta actividad nocturna, especialmente en los distritos centrales. Hay que tener en cuenta que estas zonas tienen un fuerte peso turístico y cuentan con una alta movilidad nocturna, dificultando la extracción de zonas de residencia al confundirse fácilmente con actividad nocturna. Sin embargo, los valores de correlación obtenidos indican que el número de residentes hallados por *Twitter* es una buena muestra para estudiar la distribución de la población (García-Palomares et al., 2018).

Al estudiar la relación entre el valor de los flujos hallados en la matriz obtenida con datos *Twitter* y expandida con el padrón de habitantes con los datos suministrados por la EDM, podemos observar como a nivel de municipios y distritos la correlación presenta un valor r^2 de 0,65. Sin embargo, cuando la matriz obtenida con *Twitter* se expande por los destinos usando el número de afiliados trabajadores de la Seguridad Social el ajuste se reduce a 0,35. Esta diferencia puede deberse a un mayor sesgo en el acceso a redes sociales según distintos tipos de trabajo y a que la Seguridad Social no registra a los estudiantes. En todo caso, los resultados muestran la conveniencia de expandir los datos a partir de orígenes y distribuciones de población residente.

Cuando las matrices son agregadas según grandes zonas metropolitanas y los resultados son comparados nuevamente con los obtenidos en la EDM a partir del modelo OLS, vemos como la calidad de los resultados aumenta sustancialmente, con un ajuste que casi llega al 0,85 para usuarios residentes (Figura 27). Para esa correlación es posible cartografiar los residuos (Figura 28) de manera que podamos ver en qué relaciones se producen las mayores desviaciones, ya sea por sobreestimación (mayor número de viajes a partir de datos de *Twitter* en contraste con la EDM) como por subestimación (la situación contraria). En general los resultados muestran como los datos de la matriz de *Twitter* tienen unos niveles de residuos muy bajos (lo que indica un buen ajuste respecto a los datos de la EDM), pero se puede apreciar una sobrestimación de las relaciones cuyo destino es la Almendra Central y, una subestimación las relaciones cuyo origen es la Almendra Central y cuyo destino es la periferia del municipio de Madrid, y las atracciones de los viajes internos en las zonas periféricas (Tabla 14).

Figura 27: Correlación bivariada entre valor de viajes de *Twitter* y valor de viajes de la EDM a nivel de distritos y municipio (izquierda) y zonas metropolitanas (derecha).



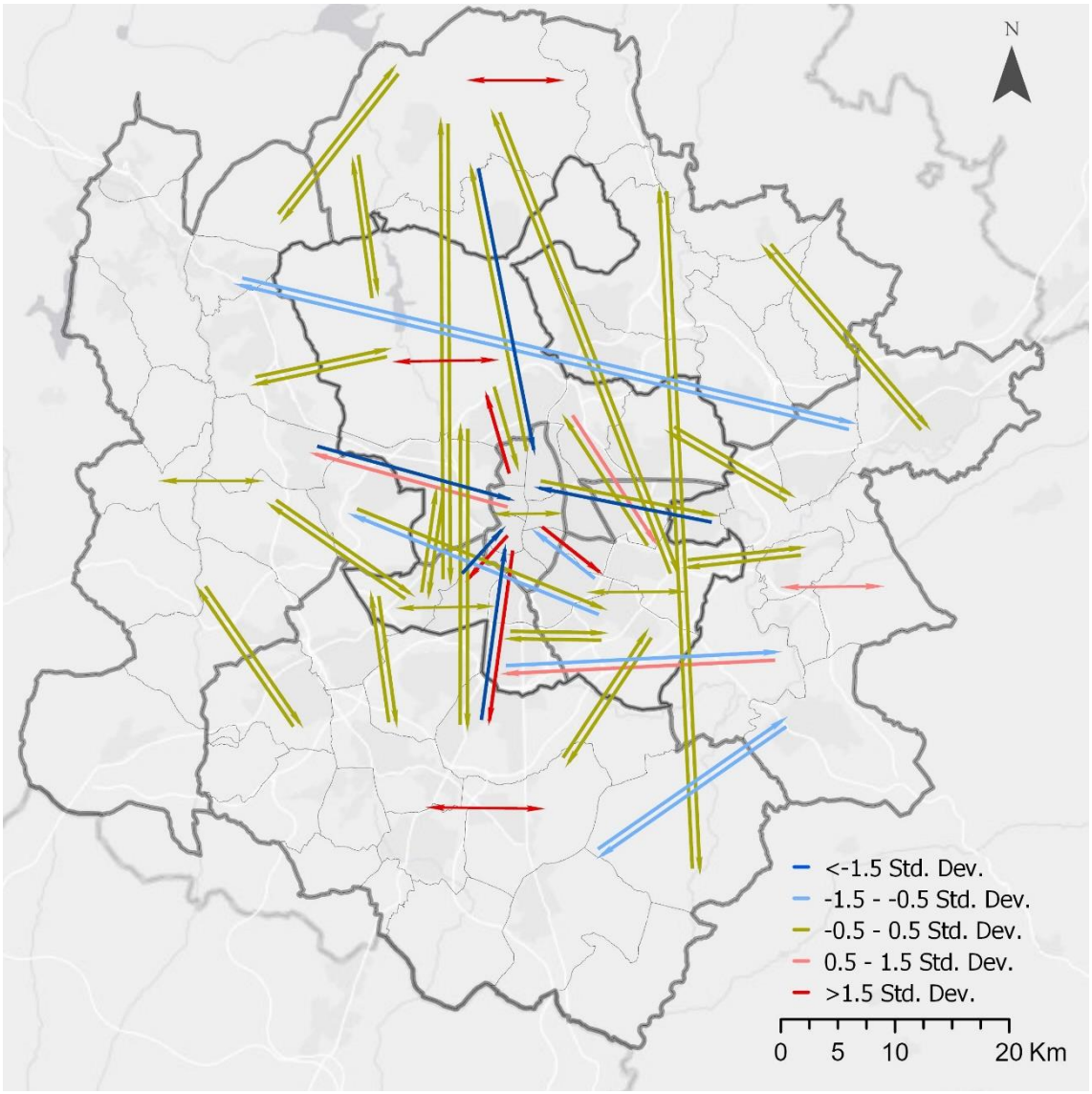
Fuente: Elaboración propia a partir de datos de *Twitter* y de la Encuesta Domiciliaria de Movilidad 2018 del Consorcio de Transportes de Madrid.

Tabla 14: Flujos de viajes a partir de datos de *Twitter* en las zonas metropolitanas de Madrid clasificados por nivel de residuos.

Totales								
	Almendra Central	Distritos Norte	Distritos Sudoeste	Distritos Sudeste	Municipios Sur	Municipios Este	Municipios Norte	Municipios Oeste
Almendra Central	68 (0,18)	1120 (2,93)	705 (1,84)	833 (2,18)	678 (1,77)	121 (0,31)	164 (0,41)	254 (0,66)
Distritos Norte	-182 (-0,47)	733 (1,92)	174 (0,45)	329 (0,86)	180 (0,47)	35 (0,09)	101 (0,26)	118 (0,31)
Distritos Sudoeste	-1037 (-2,71)	-88 (-0,22)	184 (0,48)	30 (0,07)	-8 (-0,02)	-236 (-0,61)	-192 (0,50)	-108 (-0,28)
Distritos Sudeste	-519 (-1,35)	62 (0,16)	23 (0,06)	-3 (-0,01)	-54 (-0,14)	53 (0,14)	-103 (-0,27)	-227 (-0,59)
Municipios Sur	-971 (-2,54)	-153 (-0,40)	174 (0,45)	6 (0,01)	641 (1,68)	-192 (-0,50)	-108 (-0,28)	98 (0,26)
Municipios Este	-641 (-1,68)	-52 (-0,13)	-200 (0,52)	23 (0,06)	-213 (-0,56)	369 (0,97)	-96 (-0,25)	-233 (-0,61)
Municipios Norte	-616 (-1,61)	-1 (0,00)	-164 (-0,42)	-93 (0,24)	-118 (-0,30)	-130 (-0,34)	687 (1,79)	-164 (-0,42)
Municipios Oeste	-677 (-1,77)	-48 (-0,12)	-65 (0,17)	-130 (-0,34)	122 (0,32)	-210 (-0,55)	-158 (-0,41)	103 (0,27)

Fuente: Elaboración propia a partir de datos de *Twitter* (Desviaciones estándar en paréntesis).

Figura 28: Matriz de distribución de residuos de la correlación bivariada entre valores de viajes de *Twitter* y valores de viaje de la EDM (nivel de zonas metropolitanas).



Fuente: Elaboración propia a partir de datos de *Twitter* y de la Encuesta Domiciliaria de Movilidad del año 2018 del Consorcio de Transportes de Madrid.

4.2. Visualización de caminos espacio-temporales de movilidad individual a partir de datos de *Twitter*

4.2.1. Big Data y la Geografía del Tiempo

La movilidad diaria de un individuo se estructura por la necesidad de realizar diferentes actividades (como trabajar, socializar, o comprar), que requieren estar en determinadas localizaciones durante tiempos concretos (Miller, 2005). El espacio no está completamente separado del tiempo, sino que ambos se combinan, por lo que la localización en el espacio no puede ser separada del momento temporal (Hägerstraand, 1970). La Geografía del Tiempo es una aproximación orientada a entender las actividades en esa doble componente espacio-temporal, reconociendo que la gente puede estar físicamente solo en un lugar y un tiempo en concreto (Miller, 2017).

Como se vio en el anterior caso de estudio, a medida que las ciudades han ido creciendo en población y tamaño, se han extendido en superficie y especializado internamente, con zonas dedicadas a distintas funciones. El estudio y clasificación de los usos del suelo es esencial para el planteamiento urbano. Los usos del suelo se pueden definir como el uso humano reconocido en una parcela de una ciudad, y pueden ser diferenciados por sus características físicas o por sus funciones sociales (uso residencial, comercial, ocio, etc.) (Pei et al., 2014). El número de personas en parcelas de uso del suelo residencial, de empleo o de ocio cambia según avanza el día y con ello la demanda en diferentes zonas de la ciudad, por lo que los flujos de transporte tienen comportamientos distintos en diferentes momentos del día. Por ello, una de las grandes ventajas de las nuevas fuentes de datos basadas en el *Big Data* es su capacidad de captar diferencias entre los flujos de movilidad de un área metropolitana a lo largo de un determinado periodo temporal (Gutiérrez-Puebla et al., 2019).

La ciencia de la información geográfica y las TIC están convergiendo para recrear una revolución en la que las ciencias urbanas y del transporte están pasando de ser ciencias basadas en el lugar a ciencias basadas en las personas (Miller, 2005). Los datos diarios de las actividades realizadas por muestras de individuos en un determinado periodo de tiempo han servido como fuente de datos en muchos estudios de actividades humanas espacio-temporales. Gracias a la facilidad para recoger muestras de datos creadas diariamente por las TIC de forma masiva, se abre la posibilidad de visualizar y explorar

muestras de actividades a nivel individual en un contexto espacio-temporal (J. Chen et al., 2011; Q. Huang & Wong, 2015).

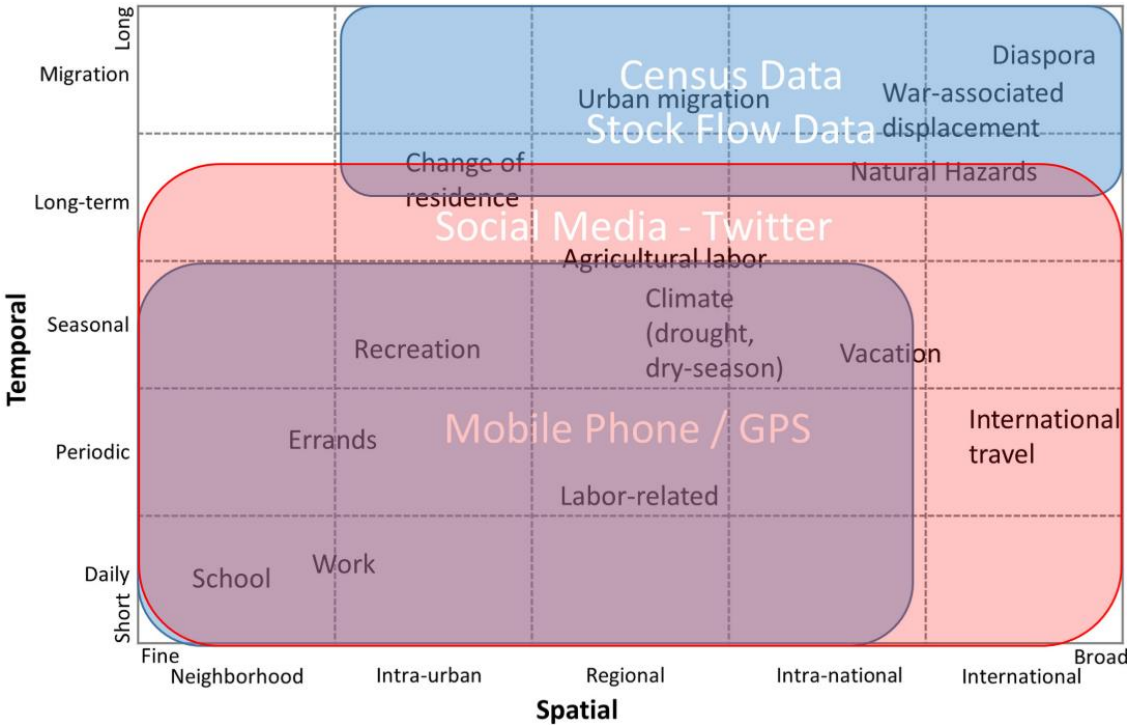
Con los avances en la computación y los SIG, es actualmente posible cartografiar los procesos que ocurren en el espacio en diferentes momentos temporales. La posibilidad de usar datos masivos para visualizar procesos espaciales en el tiempo tiene el potencial de transformar el entendimiento de la movilidad metropolitana. Juntos, la geografía del tiempo y los SIG pueden dar un ambiente analítico útil para visualizar y explorar datos de actividad a nivel individual en un contexto espacio-temporal (J. Chen et al., 2011). Como resultado, han surgido servicios y aplicaciones que publican información acerca de la actividad diaria espacial y temporal de sus usuarios (Birkin et al., 2014). La velocidad de la circulación de información a partir de estos servicios es más grande que nunca, haciendo que la gente tenga acceso sin precedentes a la información y conocimientos de características espacio-temporales de otras personas (Schwanen & Kwan, 2008).

Los datos de *Twitter* contienen información espacial y temporal precisa en forma de coordenadas donde ocurre un evento específico, permitiendo el análisis de la movilidad con un alto detalle a diferentes escalas (García-Palomares et al., 2018). La Figura 29 muestra el potencial de *Twitter* como fuente de datos para obtener datos a distintas escalas para diferentes estudios, al poder usarse en casi todas las escalas tanto espaciales como temporales. Como se mencionó en el apartado 2.3.3. de la tesis, hay que tener en cuenta que la resolución temporal de *Twitter* es menor que la de los datos de telefonía móvil. Sin embargo, aunque los datos de *Twitter* no pueden reflejar la trayectoria detallada de un usuario en el día, ofrecen localizaciones seleccionadas del individuo sobre periodos más largos de tiempos. Agrupando la información de la localización por múltiples días, los marcos de muestra temporales más largos pueden compensar la escasez temporal de la muestra en cada día, permitiendo el diseño de caminos espacio-temporales precisos (Q. Huang & Wong, 2015).

En este caso de estudio, se busca indagar en los patrones espacio-temporales de la movilidad metropolitana a partir del estudio de la movilidad individual en zonas especializadas en actividades concretas como áreas residenciales o lugares de oficinas. Para este fin se propone como metodología la visualización de caminos espacio-temporales de movilidad individual construidos a partir de datos de *Twitter*, para poder observar patrones temporales en la movilidad espacial. Además, se busca combinar estos caminos con datos del uso del suelo. Al vincular los datos de usos del suelo con la

actividad recogida en *Twitter*, es posible obtener mayor información sobre el modo en el que los usuarios de *Twitter* interactúan con el espacio en cada momento del día.

Figura 29: Clasificación de fuentes de datos a partir de escalas espacio-temporales.



Fuente: (Blanford et al., 2015).

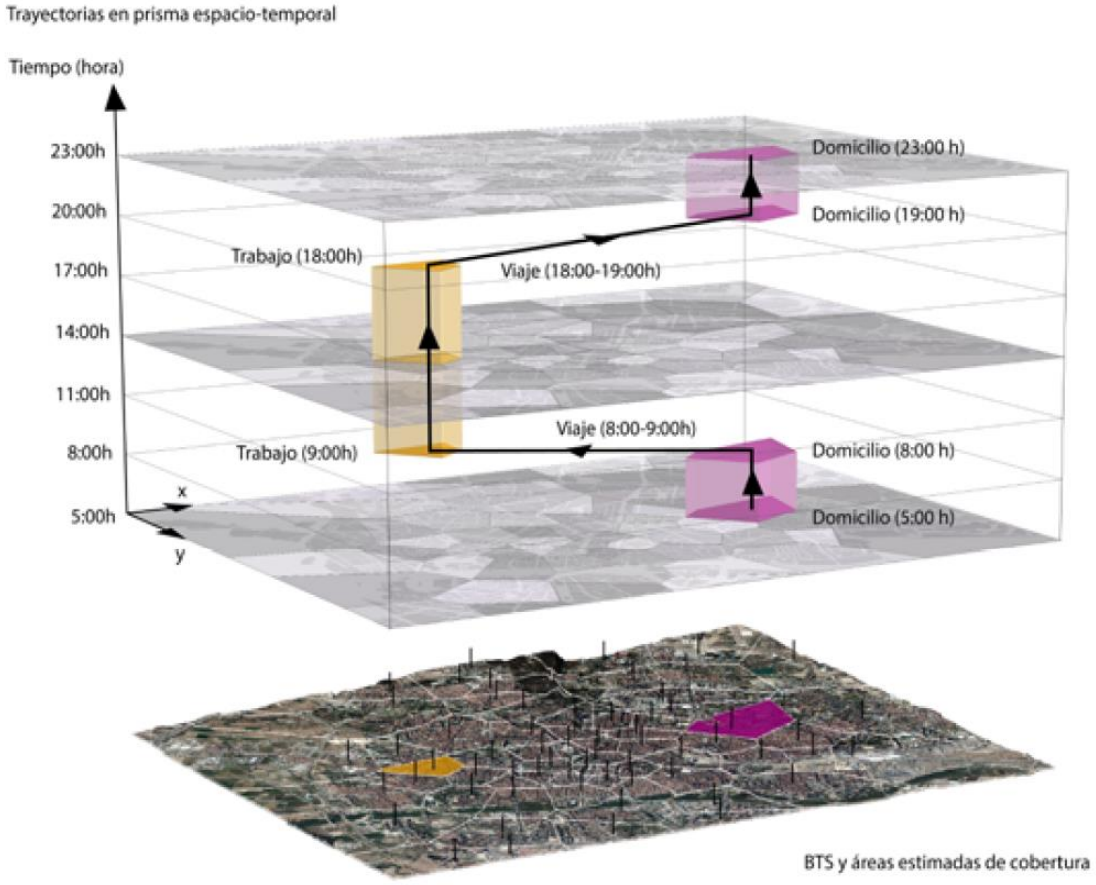
4.2.2. Metodología específica para la construcción de caminos espacio-temporales

Desde la perspectiva de la visualización, es posible representar en los SIG los datos espaciales tanto en vistas 2D como 3D dependiendo de la información suministrada. Las representaciones 2D son mejores para ilustrar relaciones espaciales precisas. Sin embargo, con esta visualización no se puede observar información basada en el tiempo (Keskin et al., 2014). Mientras, los métodos 3D en los cuales el tiempo es integrado ortogonalmente a un plano geográfico llevan de forma cualitativa a un pensamiento visual claro sobre los comportamientos humanos y los patrones geoespaciales (Neutens, Van de Weghe, Witlox, & De Maeyer, 2008). Sin embargo, hay varias dificultades técnicas y de usabilidad: la orientación del usuario en una escena visualizada (la gente puede encontrar difícil percibir la información en 3D con ángulos cambiantes), la complejidad de los datos visualizados y la falta de diseño cartográfico (Keskin, Dogru, Çelik, Doğru, & Pakdil, 2014). Por ello, en este capítulo se propone diseñar una visualización 3D para diferentes

espacios, apoyada en un mapa base de visualización 2D, para poder complementar las ventajas de cada visualización.

Para la visualización en 3D, la herramienta propuesta es el camino espacio-temporal. Este método representa el movimiento de un individuo en el espacio tridimensional a partir de una lista de puntos de control estrictamente ordenados en el tiempo. El camino se enseña gráficamente en una región espacio-temporal formada por una triada ortogonal de ejes, dos ejes x e y que definen el espacio en una llanura horizontal bidimensional, y un eje z que representa el tiempo en una dirección perpendicular. Por tanto, un objeto o punto localizado en unas coordenadas (x, y, z) muestra las coordenadas de localización del objeto en un espacio bidimensional en unas coordenada temporal z (Figura 30) (Miller, 1991, 2005). En el espacio-tiempo el individuo describe un camino en el que el lugar donde está ahora está críticamente atado al "lugar de ahora" de un tiempo anterior. El individuo no puede pasar por un punto en el espacio-tiempo más de una vez, pero tiene que estar siempre en un punto (Hägerstraand, 1970).

Figura 30: Esquema de un camino espacio-temporal.



Fuente: (Gutiérrez-Puebla et al., 2019).

Los datos de fuentes que generan entidades de puntos como *Twitter* no generan caminos espacio-temporales directamente, sino que crean una secuencia temporal de localizaciones espaciales que son usadas para construir el camino. Con estos puntos, los investigadores pueden representar caminos espacio-temporales para individuos (extrayendo los puntos por el identificador del usuario de *Twitter*), y usar estas trayectorias para visualizar la localización y el tiempo en el que se da una actividad (L. Yin, Shaw Shih-Lung, & Yu, 2011). Los datos georreferenciados suelen ser generados por usuarios de forma voluntaria, y hay que tener en cuenta que no tienen como objetivo analizar patrones de actividades. Sin embargo, aunque hay que tratar de forma cuidadosa la calidad de los datos, estos capturan algunos aspectos de las trayectorias espacio-temporales de los usuarios. Además, estos datos normalmente incluyen un gran número de usuarios para periodos relativamente largos, a coste mínimo (Q. Huang & Wong, 2015). El análisis espacio-temporal de esos datos puede revelar numerosa información oculta sobre comportamientos humanos en el espacio y tiempo, y sobre relaciones que afectan estos comportamientos con otras variables que afectan la movilidad (Keskin et al., 2014).

Para mejorar la resolución temporal de los datos de *Twitter* se ha trabajado agregando los *tweets* de múltiples días laborables, con el fin de obtener localizaciones concretas suficientes en una secuencia de 24 horas y diseñar así los caminos espacio-temporales diarios de cada usuario (Q. Huang & Wong, 2015). Estos *tweets* ya están previamente enriquecidos con datos del uso del suelo del Catastro. Al trabajar con *tweets* en días laborables se pueden hallar comportamientos regulares de movilidad, mientras que la movilidad urbana en los fines de semana es más errática. A partir de las fechas de cada *tweet*, y trabajando con un total de 516 días laborables, se ha podido calcular el número de días que cada usuario ha twitteado en un lugar y hora determinada.

Por cada usuario, se agregó el número de hora y el identificador de la zona de transporte de cada uno de sus *tweets*, con el objetivo de extraer en que zonas ha publicado cada usuario un mayor número de *tweets* en un número determinado de hora. Sin embargo, un individuo puede tener más de una localización visitada regularmente o múltiples trayectorias de viaje. Los puntos que están relativamente lejos de otros puntos tanto espacialmente como temporalmente pueden ser resultados de actividades aleatorias. En cambio, puntos cercanos espacio-temporalmente reflejan actividades regulares (Q. Huang & Wong, 2015). En casos en los que un usuario ha publicado mensajes desde más de una

zona en una misma franja horaria, se seleccionó la zona de transporte donde se *twitteó* un mayor número de días en dicha hora. Si la frecuencia máxima de días en una determinada hora coincide en más de una zona, se seleccionó la zona teniendo como referencia las zonas en las que se había escrito con mayor frecuencia en la hora anterior y posterior a la tratada.

Esta metodología de trabajar con los datos agrupados para el conjunto de días permite conocer patrones de actividad regular de los usuarios, pero pueden incluir actividades irregulares, introduciendo ruido o incertidumbre (Q. Huang & Wong, 2015). Para calcular caminos espacio-temporales de recorrido recurrente, se han definido estancias en cada una de las franjas horarias para las localizaciones de zonas de transporte donde el usuario ha *twitteado* durante un mínimo de 3 días laborables. De esta forma se tratan de minimizar los puntos con potencial actividad aleatoria. Entonces, se simplificó la base de datos para que solo hubiese un punto registrado por usuario, lugar, y zona de transporte. Como resultado, se han obtenido 18.923 puntos con los que se puede simular el camino espacio-temporal de 2.706 usuarios en el Área Metropolitana de Madrid. Mediante técnicas estadísticas de resumen de datos es posible cartografiar la distribución de los usuarios en los diferentes distritos y municipios del área de estudio tanto por hora como por uso del suelo.

A continuación, se seleccionaron cuatro áreas de estudio con una orientación a un uso de suelo concreto en las horas más propicias para el desarrollo de determinadas actividades. Estas cuatro zonas de estudio fueron las de Puente de Vallecas, Nuevos Ministerios-AZCA, Ciudad Universitaria, y Parque de Retiro. Puede verse más información de las mismas en el capítulo 3.1.1. de la tesis doctoral y su localización en la Figura 16. El siguiente paso fue seleccionar de la base de datos los usuarios que tengan un punto dentro de cada franja temporal definida en cada zona de estudio. La Tabla 15 presenta el número de usuarios válidos para los que se han obtenido caminos espacio-temporales en las distintas zonas de estudio. Finalmente, para la representación de los recorridos espacio-temporales en 3D, se ha otorgado a cada punto un valor de altura igual al número de hora, multiplicado por un factor de exageración de 200. Con estos valores, se han construido capas de líneas para representar los caminos espacio-temporales.

Al haber sido enriquecidos previamente con datos de uso del suelo del Catastro, cada punto de los caminos espacio-temporales diseñados tiene información del uso del suelo en ese momento. A partir de estos puntos, se ha calculado el porcentaje de usuarios en

días laborables tanto por hora como por uso del suelo sobre el total de la muestra. De esta forma, es posible visualizar los usos del suelo principales en las cuatro zonas de estudio a lo largo del día.

Tabla 15: Número de usuarios en cada zona de estudio.

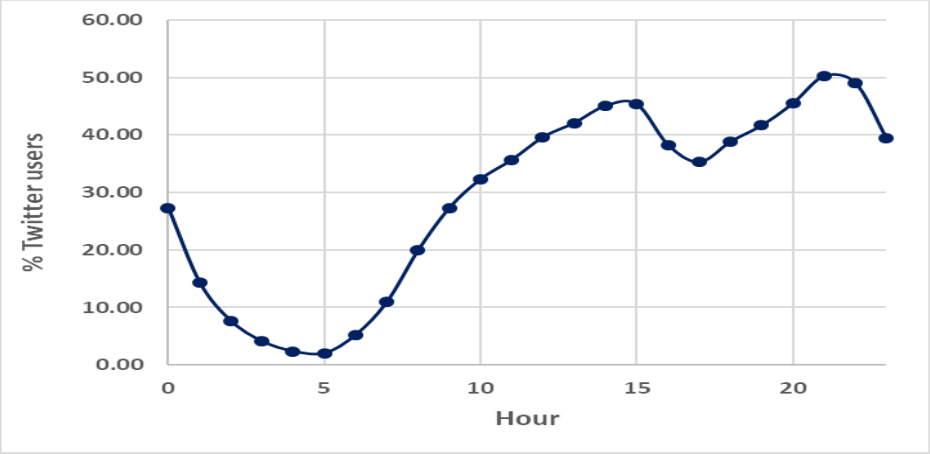
Zona	**Horario**	**Tipo**	**Número usuarios validos**
Puente de Vallecas	Noche (21 a 8 horas)	Residencial	27
Nuevos Ministerios-AZCA	Mañana (8 a 15 horas)	Trabajo	19
Ciudad Universitaria	Mañana (8 a 15 horas)	Estudios	30
Parque del Retiro	Tarde (16 a 21 horas)	Ocio	39

Fuente: Elaboración propia a partir de datos de *Twitter*.

4.2.3. Distribución de los usuarios de Twitter en el tiempo

Tomando la muestra de 2.706 usuarios de *Twitter*, se puede observar un continuo aumento del porcentaje de usuarios a lo largo del día hasta las 15 horas de la tarde, momento que coincide con el regreso de muchos trabajadores o estudiantes a sus hogares. A continuación, se da un descenso de usuarios hasta las 18 horas. En esta hora de nuevo crece el número de usuarios hasta alcanzar su punto máximo a las 21 horas. A partir de esta hora el número de usuarios crece de forma continua (Figura 31 y Tabla 16).

Figura 31: Porcentaje de usuarios de *Twitter* durante el día.



Fuente: Elaboración propia a partir de datos de *Twitter*.

Tabla 16: Usuarios de *Twitter* por hora.

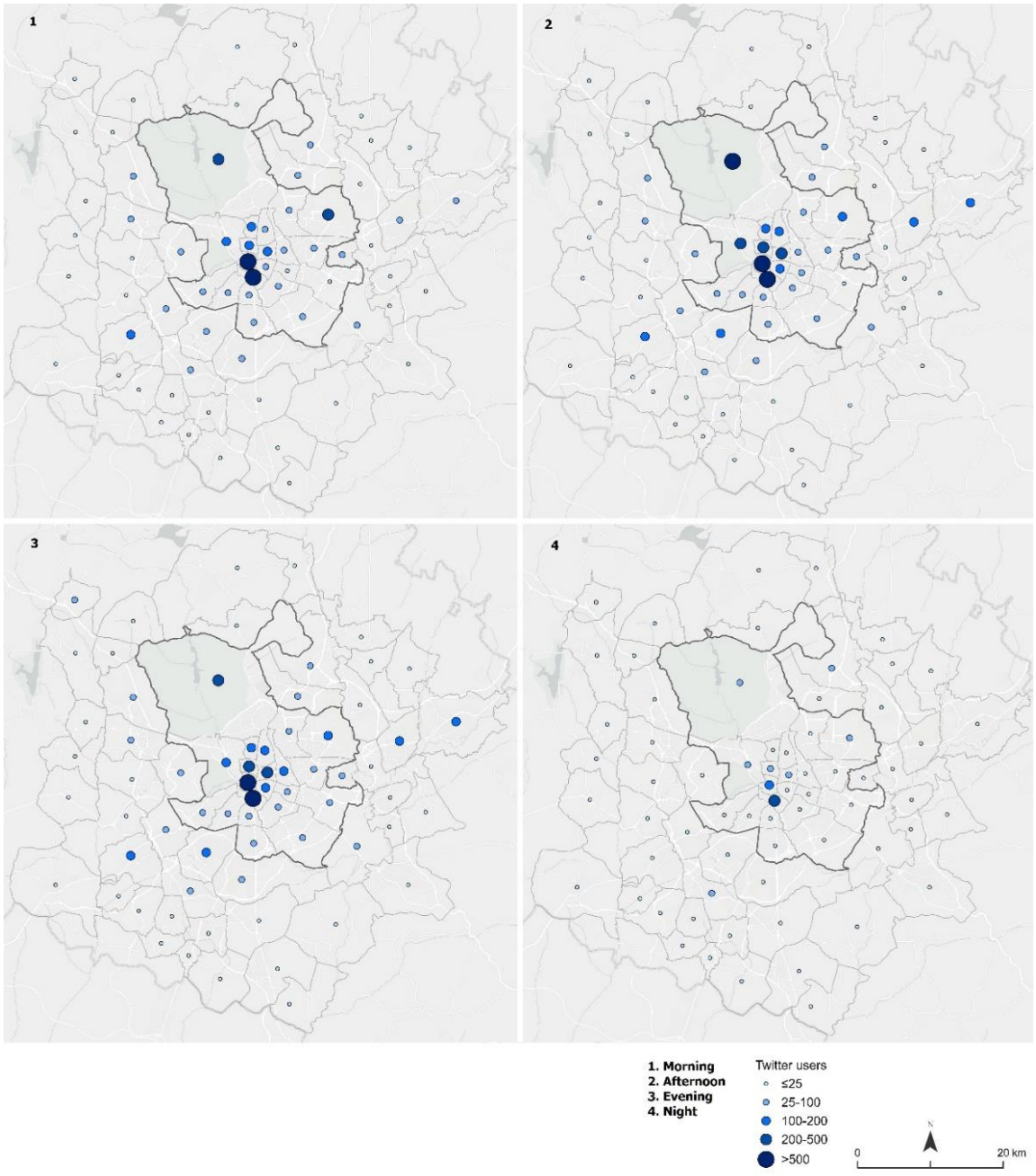
Hora	**Total usuarios *Twitter***	**% usuarios *Twitter***
0	739	27,31
1	389	14,38
2	207	7,65
3	111	4,10
4	64	2,37
5	55	2,03
6	141	5,21
7	296	10,94
8	539	19,92
9	739	27,31
10	874	32,30
11	965	35,66
12	1.072	39,62
13	1.139	42,09
14	1.219	45,05
15	1.227	45,34
16	1.035	38,25
17	954	35,25
18	1.050	38,80
19	1.127	41,65
20	1.231	45,49
21	1.359	50,22
22	1.327	49,04
23	1.065	39,36

Fuente: Elaboración propia a partir de datos de *Twitter*.

La Figura 32 muestra la distribución de usuarios de *Twitter* en los distritos y municipios del Área Metropolitana de Madrid en diferentes momentos del día. Se puede apreciar como la actividad se concentra en la Almendra Central de la ciudad de Madrid durante todo el día, especialmente por la mañana. A mediodía se puede apreciar un aumento del número de usuarios en el norte del municipio de Madrid y en la zona este del área metropolitana, mientras que los municipios del sur del Área Metropolitana ganan usuarios durante la tarde. Se observa por tanto un importante número de usuarios que trabajan en

el centro de la ciudad por la mañana y luego regresan a sus hogares en los municipios periféricos del sur y del este por la tarde.

Figura 32: Número de usuarios de *Twitter* en diferentes momentos del día.

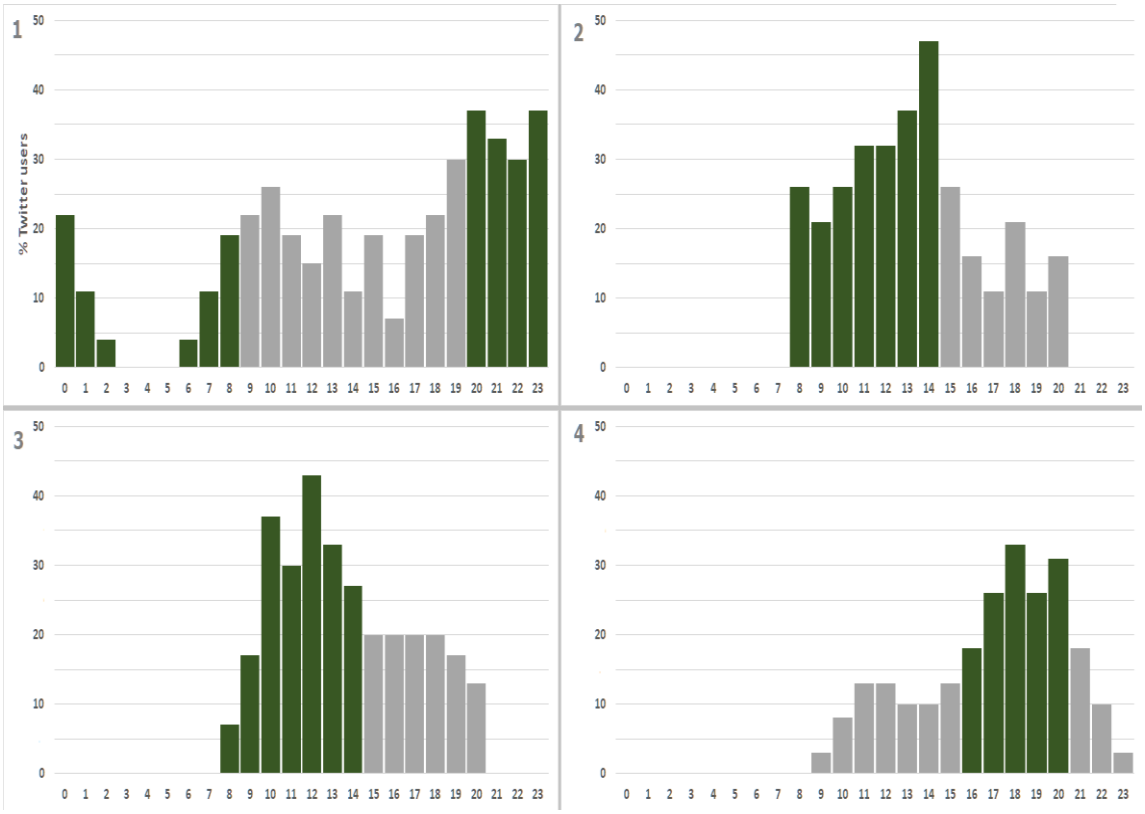


Fuente: Elaboración propia a partir de datos de *Twitter*.

Los resultados obtenidos muestran las actividades de los usuarios en distintos momentos del día acorde con los usos del suelo principales en cada una de las zonas de estudio seleccionadas (a partir de las muestras de usuarios detalladas en la Tabla 14). La zona residencial de Puente de Vallecas muestra una disminución gradual del porcentaje de

usuarios durante la mañana, y un aumento de la actividad a lo largo de la tarde, hasta llegar al horario de noche donde se concentran los mayores porcentajes de usuarios. Esta es una pauta característica del carácter residencial de la zona. En el área de oficinas de Nuevos Ministerios-AZCA se aprecia una mayor concentración de usuarios a lo largo de la mañana (con un pico determinado en las 14 horas, final del horario laboral por la mañana) y un fuerte descenso de usuarios por la tarde. El caso de Ciudad Universitaria es relativamente similar, aunque el porcentaje de usuarios aumenta de forma más brusca a primeras horas de la mañana y desciende más lentamente por la tarde. En el Parque del Retiro, el porcentaje de usuarios va en aumento a lo largo del día, con predominancia de usuarios en el horario de tarde. Cuando llega la noche, este porcentaje de usuarios empieza a disminuir prolongadamente (Figura 33).

Figura 33: Distribución de usuarios por hora en Puente de Vallecas (1), Nuevos Ministerios-AZCA (2), Ciudad Universitaria (3), y Parque del Retiro (4).

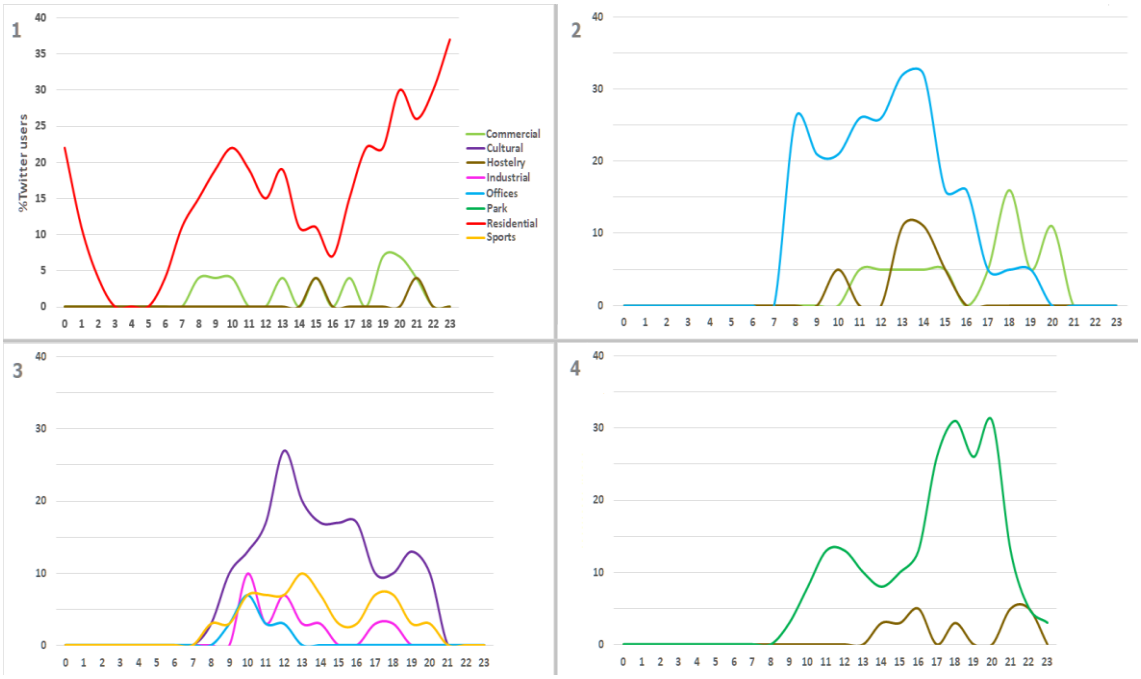


Fuente: Elaboración propia a partir de datos de *Twitter*.

La distribución temporal de usuarios basada en los usos del suelo corrobora la situación observada en las zonas de estudio. En la zona de Puente de Vallecas, el principal uso del suelo, el residencial, aumenta continuamente a lo largo del día a partir de primera hora de la tarde hasta llegar a su máximo por la noche. Se puede observar en menor grado un

constante uso comercial a lo largo del día, y un uso de hostelería en las horas de almuerzo y cena. Nuevos Ministerios-AZCA presenta un uso del suelo principal de oficinas en horario de mañana y que va descendiendo por la tarde, cuando a su vez, el uso comercial pasa a ser la principal actividad de los usuarios de la zona. También se pueden apreciar dos picos de hostelería a la hora del desayuno y almuerzo. El uso cultural es predominante en Ciudad Universitaria durante todo el día, especialmente durante la mañana y las primeras horas de la tarde. Se puede apreciar usos del suelo de oficinas e industrial por la mañana, y uso de los espacios deportivos a últimas horas de la mañana y durante la tarde. Finalmente, el Parque del Retiro presenta sus mayores porcentajes de usos del suelo en horario de tarde, a la vez que paralelamente se da uso del suelo de hostelería (Figura 34).

Figura 34: Distribución de usuarios por usos de suelo y hora en Puente de Vallecas (1), Nuevos Ministerios-AZCA (2), Ciudad Universitaria (3), y Parque del Retiro (4).



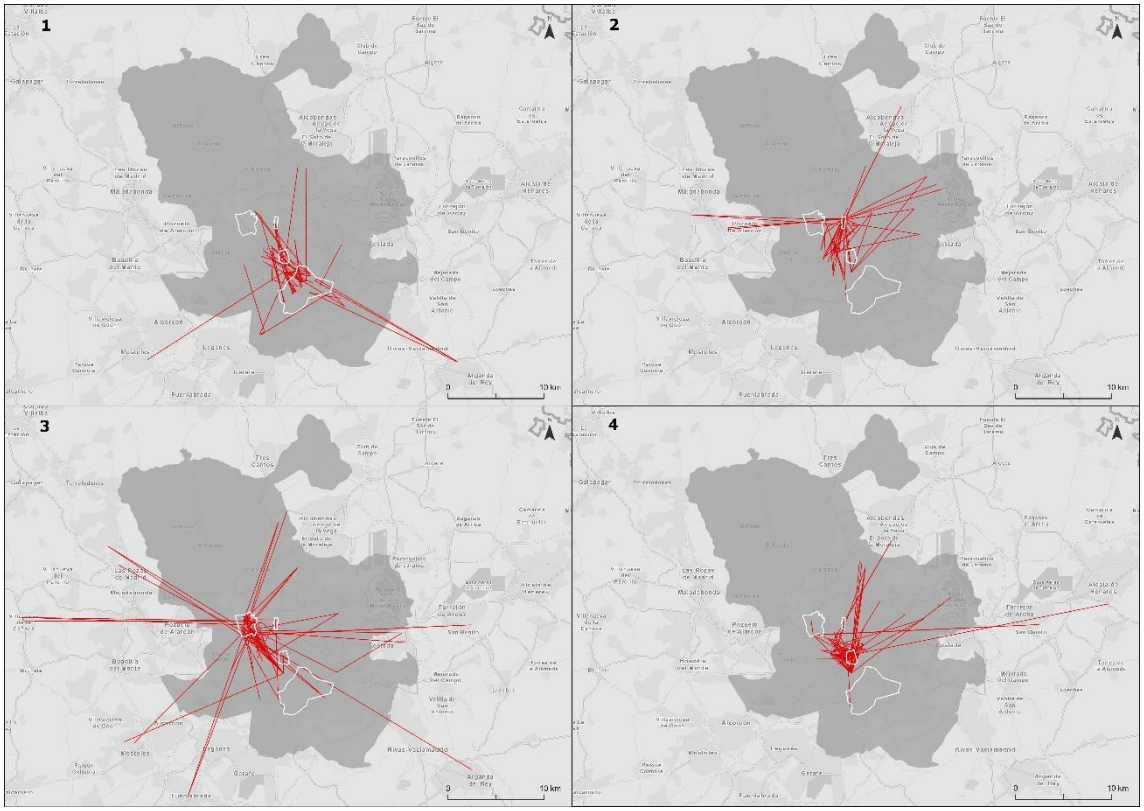
Fuente: Elaboración propia a partir de datos de *Twitter*.

4.2.4. Visualización de trayectorias individuales de movilidad en el espacio-tiempo

La visualización de los caminos espacio-temporales en 2D muestra los destinos de los desplazamientos para los residentes en el espacio residencial de Puente de Vallecas y los orígenes de los trabajadores en la zona de oficinas de AZCA, los estudiantes de Ciudad Universitaria o los usuarios del parque del Retiro (Figura 35). Mientras, la visualización

en 3D permite visualizar los movimientos de estos usuarios a lo largo del día en las zonas de estudio (Figura 36).

Figura 35: Caminos espacio-temporales a lo largo del día (2D) en Puente de Vallecas (1), Nuevos Ministerios-AZCA (2), Ciudad Universitaria (3), y Parque del Retiro (4).



Fuente: Elaboración propia a partir de datos de *Twitter*.

Combinando ambas visualizaciones, se puede apreciar que los residentes detectados en Puente de Vallecas salen del barrio preferentemente a primera hora de la mañana y se desplazan principalmente a la cercana estación de trenes de Atocha, desde donde se mueven a otros puntos de la ciudad (principalmente el centro o zonas del norte de la ciudad que concentran la oferta de trabajo). También se observan algunos movimientos de regreso y salida entre las 14 y las 16 de la tarde, usuarios que vuelven a la residencia a almorzar y vuelven al trabajo, y un regreso al barrio a partir de las 20 horas. Paralelamente se visualiza un movimiento interno de usuarios que se desplazan a lo largo del día por la zona, mostrando el carácter dinámico de un distrito de población obrera que también cuenta con actividad de trabajo y ocio internos.

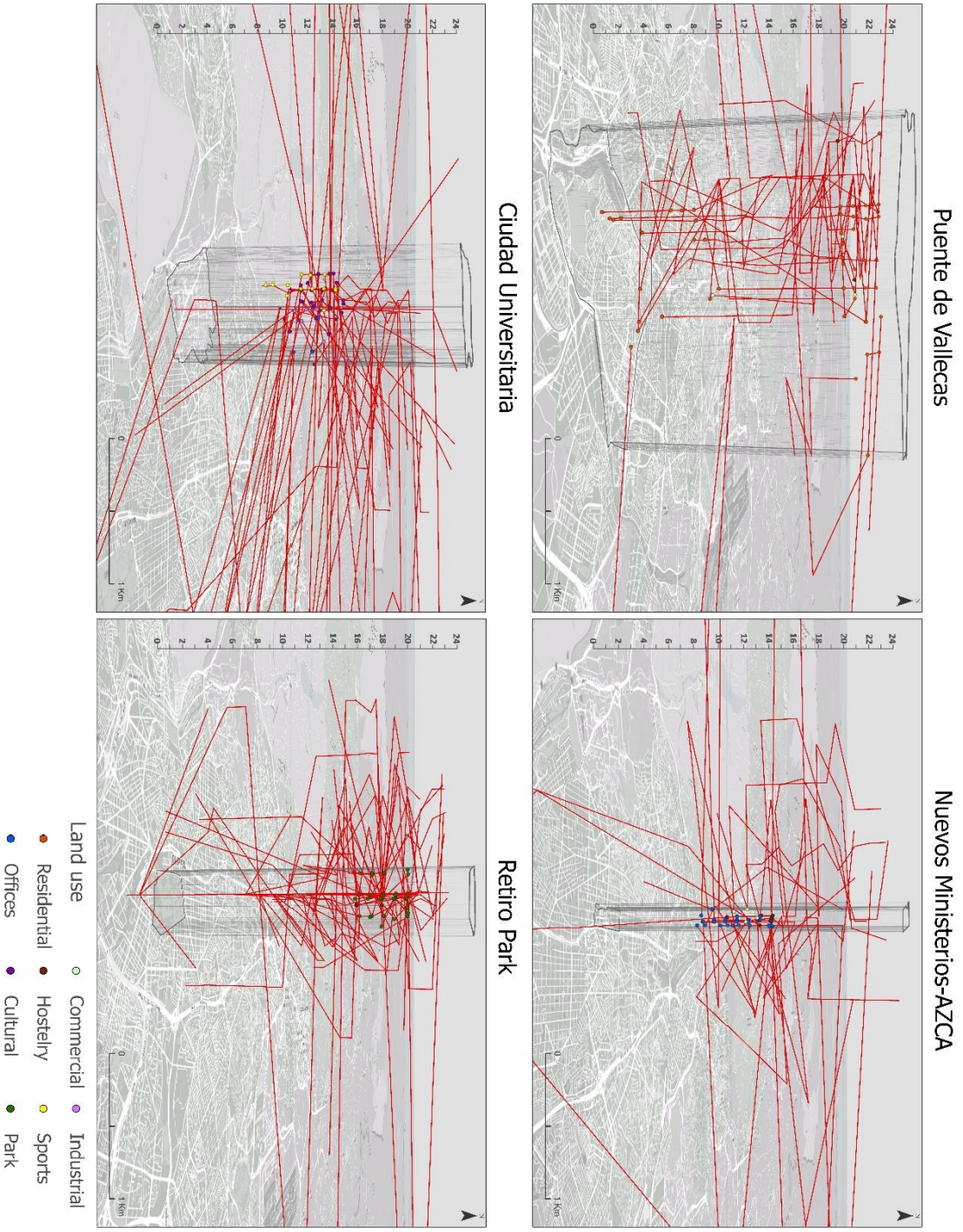
Mientras, la zona de oficinas de Nuevos Ministerios-AZCA recibe usuarios en horario de mañana, que provienen principalmente de las estaciones de tren de la ciudad (usuarios del área metropolitana que viajan a la zona en tren) y también del este de la ciudad o de los

municipios colindantes del norte o del oeste (caracterizados por una población empresarial con niveles altos de renta). Igualmente, se aprecia una amplia diversificación de los caminos al resto de la ciudad a lo largo del día, principalmente por la tarde (usuarios que retornan a su residencia).

Una situación similar aparece en la Ciudad Universitaria, donde destaca un número importante de caminos a partir de las estaciones de tren, que van a la universidad en horario de mañana y salen de la universidad constantemente durante la tarde. En este caso se observa un flujo de caminos importante que proviene principalmente de las zonas residenciales del oeste y del sur del área metropolitana, pero que cuenta con un área de influencia mayor que la zona de oficinas. Además, las horas del día en la que los estudiantes llegan a Ciudad Universitaria están más concentradas, principalmente por la mañana y a primera hora de la tarde.

En el Parque del Retiro se llevan a cabo movimientos durante todo el día, provenientes principalmente de espacios próximos en norte y el este de la ciudad y con una mayor frecuencia en horario de tarde. La visualización en 2D muestra un área de influencia de sus visitantes mucho más pequeña en comparación con las dos zonas anteriores de trabajo y estudios, e indica un desplazamiento de ciudadanos principalmente del propio municipio de Madrid debido al propio carácter temático y recreacional del parque.

Figura 36: Caminos espacio-temporales a lo largo del día (3D).



Fuente: Elaboración propia a partir de datos de *Twitter*.

4.3. Estudio de la movilidad universitaria a partir de datos de *Twitter*

4.3.1. Interés de la movilidad universitaria

Los campus universitarios son frecuentados por población de diferentes edades y fondos socioeconómicos, poseen horarios irregulares, y generan un movimiento constante de gente durante el día (Miralles-Guasch & Domene, 2010; Soria-Lara, Marquet, et al., 2017). Las universidades son centros de oferta o atracción de viajes que requieren de una infraestructura necesaria para soportar grandes volúmenes de viajeros (Whalen et al., 2013). Las áreas metropolitanas suelen albergar una o más universidades de gran tamaño. Cuanto mayor sea el número de habitantes de una ciudad, más fácil será encontrar campus de diferentes universidades. El transporte a los campus está afectado por múltiples causas, que se pueden resumir en factores de localización espacial (localización de las universidades, distancia entre la residencia y la universidad, abundancia de transporte público, densidad urbana), factores socioeconómicos de la población universitaria (nivel de renta, edad), y factores de comportamiento social (factores que influyen en el estilo de vida de la población universitaria como grupo colectivo: duración de viajes diarios, número de días a la semana en los que se viaja a la universidad, etc.) (Soria-Lara, Marquet, et al., 2017).

El análisis de la movilidad universitaria es importante para los gestores de transporte y para las propias universidades, para la planificación de transportes, horarios, residencias o servicios públicos (Soria-Lara, Miralles-Guasch, et al., 2017). Para las universidades, promover el transporte sostenible es una herramienta para demostrar que hay un valor añadido en la educación universitaria (Davison et al., 2015). Las universidades, como espacios de atracción de viajes, implementan estrategias para reducir la dependencia de vehículos privados y aumentar el uso de transportes alternativos para reducir la demanda de sitios de aparcamiento y los impactos medioambientales de los desplazamientos (Shannon et al., 2006). En el campo de la exclusión social urbana, la segregación se entiende cómo la restricción de la presencia de los individuos a las acciones diarias en la ciudad (Netto et al., 2015). Partiendo del hecho de que las dificultades en el transporte pueden facilitar la exclusión social a partir de la falta de acceso a oportunidades y servicios, se forma una barrera que impide el acceso a estudios superiores. Comprender la movilidad universitaria puede ayudar a crear un modelo sostenible medioambiental,

que disminuya la dependencia del automóvil, ayude a disminuir la exclusión social urbana, y permita el acceso a oportunidades o instituciones.

Tradicionalmente, los estudios relacionados sobre la movilidad universitaria se han realizado a partir de encuestas como fuente de datos. Sin embargo, la irrupción del *Big Data* facilita el acceso a datos que permiten complementar la información obtenida en encuestas tradicionales de forma rápida, barata, con muestras grandes, y con alto detalle espacio-temporal. Entre las nuevas fuentes de datos, las redes sociales son una de las herramientas con las que se puede conseguir datos masivos y de alta resolución espacio-temporal de población universitaria potencial a tiempo casi real. Un aspecto a tener en cuenta es el sesgo producido por el uso diferencial de las redes sociales según franjas de edad. Casi el 40% de usuarios de *Twitter* tienen entre 18 y 30 años. Sin embargo, este sesgo es una ventaja a la hora de estudiar la movilidad universitaria, ya que justo el grupo de edad más representado es en el que se encuentran los estudiantes universitarios.

En este caso de estudio se propone detectar población universitaria a partir de datos obtenidos por *Twitter*, y valorar hasta qué punto *Twitter* puede ser una herramienta de utilidad en cuestiones relacionadas con la atracción de viajes que causan los campus universitarios. Para ello, se ha trabajado con datos de *Twitter* correspondiente a los cursos escolares 2016/17 y 2017/18 en el área metropolitana de Madrid, tratando de establecer las áreas de influencia de las distintas universidades madrileñas y estudiar las diferencias según el tipo de universidad. A continuación, se busca estudiar factores de la movilidad a los campus analizando los datos de tiempos de viaje según el modo de transporte, el tipo de universidad y el nivel adquisitivo de la zona donde residan los estudiantes. Finalmente, se pretende crear un modelo gravitacional con el que comparar una asignación teórica de usuarios a las universidades a partir de datos oficiales, y compararlo con los resultados de la asignación de usuarios identificada a partir de *Twitter*. En este caso, se parte de la premisa de que los datos de *Twitter* se aproximan más a la realidad ya que el modelo teórico que únicamente considera la proximidad y el tamaño de los campus, lo que permite mostrar que existen otros factores más allá de estos dos en la elección de los centros universitarios.

4.3.2. Metodología específica para el cálculo de áreas de influencia y del modelo de asignación de población universitaria

Una vez realizados los procesos de limpieza y filtrado de usuarios, el primer paso fue detectar usuarios de los campus universitarios, en su mayoría estudiantes. Para ello, se clasificaron los *tweets* en función de si fueron publicados en el horario de día (8 de la mañana a 8 de la tarde) o de noche (el resto). A continuación, se construyó una capa de polígonos de las facultades universitarias de la Comunidad de Madrid a partir del Nomenclátor oficial y Callejero de la Comunidad de Madrid, y se corrigieron posibles errores tomando como referencia información del catastro y del callejero *OpenStreetMap*. Después, se realizó un buffer para seleccionar *tweets* publicados durante el día a unas determinadas distancias de los edificios universitarios. Para facultades fuertemente integradas en el tejido urbano de la ciudad, se empleó un buffer de 20 metros, para facultades dentro del tejido urbano, pero con cierta separación respecto a otros edificios y usos del suelo, la distancia realizada fue de 50 metros, y para facultades localizadas en áreas específicas orientadas a albergar solamente campus universitarios, se buscaron los *tweets* a una distancia máxima de 100 metros. Como resultado, se obtuvieron 15.380 mensajes realizados por 5.296 personas (el 3% de usuarios respecto a la base inicial). A cada usuario se le asignó la universidad y campus desde la cual publicó un mayor número de mensajes.

A continuación, se expandió la base de datos a partir de la búsqueda de los últimos 3.200 *tweets* de cada uno de los 5.296 posibles usuarios encontrados previamente, con el objetivo de aumentar la precisión espacial y temporal de los movimientos individuales de cada uno de estos usuarios. Como resultado, la base de datos alcanzó un total de 18.951 mensajes. Tras dicha expansión, se procedió a buscar el lugar de residencia de cada uno de estos usuarios, siguiendo la metodología explicada en el primer caso de estudio correspondiente al apartado 4.1. de la tesis. Para cartografiar las áreas de influencia de cada una de las universidades de la Comunidad de Madrid, se ha usado el lugar de residencia de los 2.912 estudiantes encontrados según municipios y distritos. Cada uno de estos usuarios tiene un distrito o municipio asignado como residencia, además de una universidad y campus de destino. Con esta información se ha obtenido una matriz residencia-campus. La matriz cuenta con 70 orígenes (21 distritos de la ciudad de Madrid y 49 municipios del área metropolitana), y 34 destinos (campus universitarios). Sin embargo, del total de 2.380 relaciones, 680 tenían datos de usuarios. Esta matriz ha sido

expandida a partir de los datos oficiales de número de estudiantes según campus usando la siguiente fórmula:

$$E_{ij}^{e} = E_{ij}.\frac{p_j}{\tilde{p}_j}, \forall j \in N,$$

donde $\frac{p_j}{\tilde{p}_j}$ son los pesos calculados para cada campus N, basándose en el número total de estudiantes por campus según las fuentes oficiales p_j y la muestra de usuarios de *Twitter* por campus $\tilde{p}_j$. Estos pesos son multiplicados por el valor de los viajes identificados con *Twitter* en cada flujo E_{ij}, siendo el resultado flujos expandidos E_{ij}^{e}.

El cálculo de distancias medias recorridas se elaboró a partir de la matriz obtenida en el paso anterior. Con los datos de *Twitter* no es posible identificar el modo de transporte utilizado desde el lugar de residencia al campus. Sin embargo, se puede estimar el tiempo de viaje desde cada municipio al campus tanto en transporte público como privado. A partir de esos datos se calcularon tiempos ponderados por el número de usuarios residentes en el municipio usando como fuente de datos los tiempos de viaje obtenidos a partir de ficheros *GTFS* para el cálculo de tiempos ponderados a partir de tiempos de transporte público, y datos de transporte privado *TomTom* para la obtención de tiempos ponderados mediante tiempos de transporte privado. El tiempo ponderado para cada universidad se calculó a partir de la siguiente fórmula:

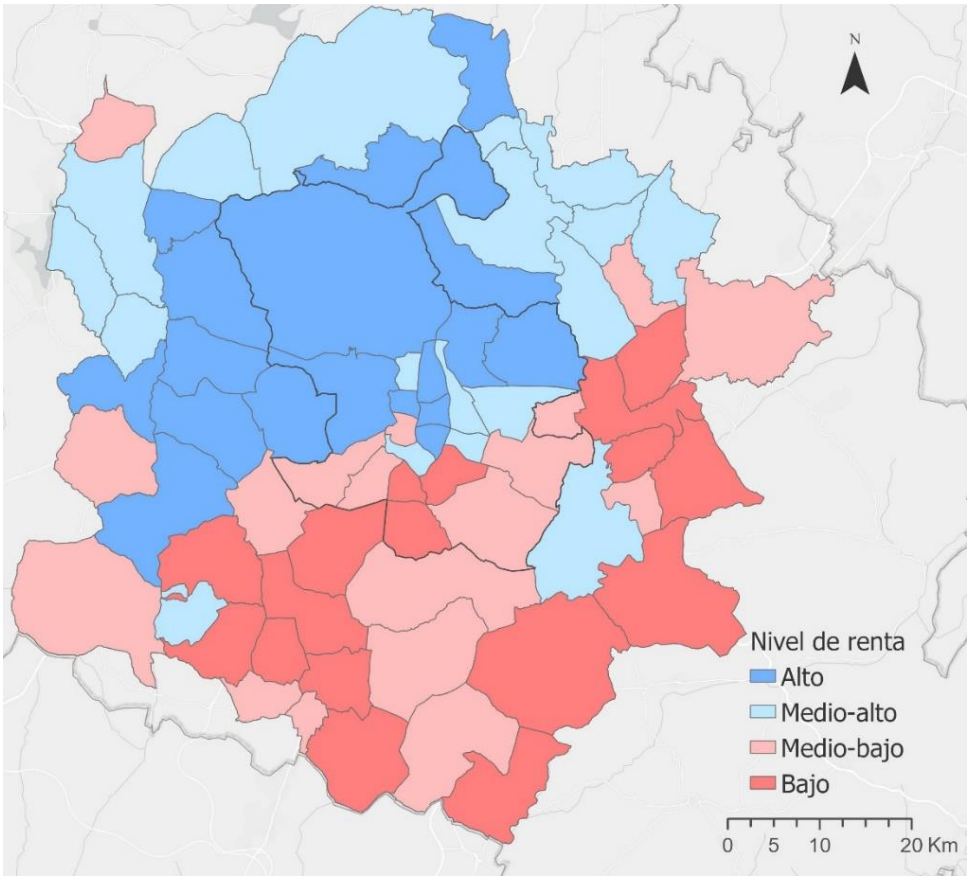
$$T_u = \sum_{i \in N}(\sum T_{ic} \frac{\tilde{p}_c}{\sum \tilde{p}_u}) \frac{\tilde{p}_i}{\sum \tilde{p}_i}, \forall u \in U,$$

donde T_u es la media del tiempo ponderado de cada universidad *u*, *N* es el conjunto de todos los municipios y distritos del área de estudio, T_{ic} es el tiempo registrado de transporte público o privado una ruta entre un municipio *i* y un campus *c*, $\tilde{p}_i$ es la muestra de usuarios de *Twitter* por municipio, y $\frac{\tilde{p}_c}{\sum \tilde{p}_u}$ es el valor de ponderación de un campus a partir de $\tilde{p}_c$ (el número de usuarios de *Twitter* obtenidos en un campus *c*) y $\tilde{p}_u$ (el número de usuarios de *Twitter* de una universidad *u*).

Para el análisis de los tiempos de acceso a la universidad según el nivel de renta del lugar de origen, se han agrupado los distritos y municipios por cuartiles según su nivel de renta, clasificándolos según su cuartil en municipios/distritos de renta alta, media-alta, media-baja, y baja (Figura 37). Se ha calculado para cada uno de estos grupos el porcentaje de usuarios cuya residencia ha sido estimada, y la media de tiempos de viaje a partir de los datos obtenidos previamente a nivel de distrito o municipio. El porcentaje de usuarios

cuya residencia ha sido estimada mediante datos de *Twitter* corresponde al 0,62% respecto al total de la población joven entre 18 y 25 años según el censo (Tabla 17).

Figura 37: Municipios y distritos del Área Metropolitana de Madrid según nivel de renta por cuartiles.



Fuente: Elaboración propia a partir de los datos de 2016 del Instituto de Estadística de la Comunidad de Madrid y datos de 2015 del Urban Audit del Ayuntamiento de Madrid.

Tabla 17: Porcentaje de universitarios encontrados en *Twitter* respecto a datos censales por cuartil de renta.

Nivel de renta	**% Universitarios *Twitter* sobre población total 18-25 años**
Alto	0,78
Medio-alto	0,42
Medio-bajo	0,95
Bajo	0,27
Total	0,62

Fuente: Elaboración propia a partir de datos de *Twitter*.

Con el objetivo de comparar los datos obtenidos de *Twitter* con un modelo basado en datos oficiales, se ha elaborado un modelo gravitacional de *Huff* que permite obtener una asignación de población universitaria a cada campus a partir de plazas ofertadas y tiempos de transporte de acceso a los mismos, a partir de la siguiente fórmula:

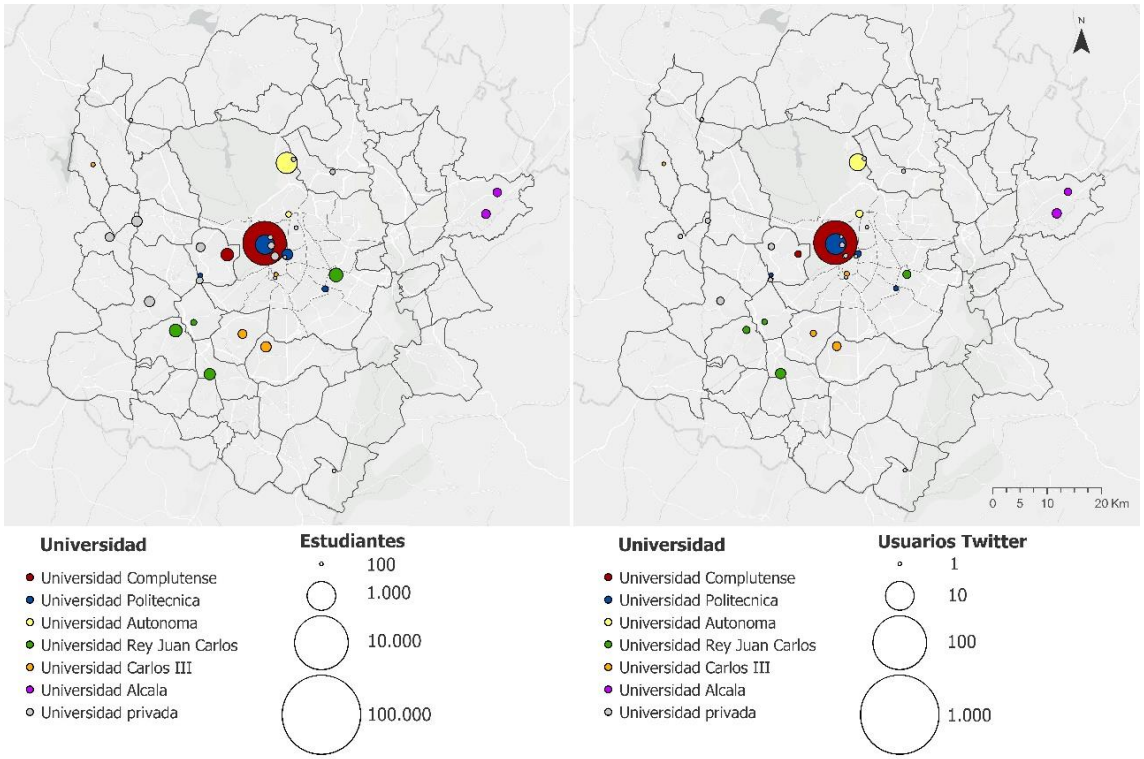
$$N_{iu} = \frac{O_u \cdot \frac{1}{T_u}}{\sum_{O \in u} O_u \cdot \frac{1}{T_u}} . D_i, \forall u \in U,$$

donde el número de universitarios N_{iu} del municipio i asignado al campus u se calcula a partir de la oferta de plazas de cada universidad O_u, la inversa del tiempo desde cada municipio o distrito a cada universidad T_u, y el producto del número de habitantes (entre 18 y 25 años) por municipio *i* obtenidos mediante datos censales. Para la oferta de plazas por universidad se utilizaron datos oficiales del Ministerio de Educación, Cultura y Deporte. Una vez obtenido el número estimado de personas asignados desde cada municipio o distrito a cada universidad por el modelo de *Huff*, se estableció una relación con el número de usuarios residentes obtenidos en *Twitter*, pudiendo así conocer en que municipios y distritos se han obtenido en *Twitter* un mayor número de usuarios por universidad respecto al modelo gravitacional.

4.3.3. Lugares de residencia de los usuarios de los campus y áreas de influencia de las universidades

Tras los procesos de limpieza y filtrado de datos, la base final de *tweets* alberga 5.296 usuarios, cada uno con una universidad y campus asignado. El coeficiente de correlación (r^2) entre el número de usuarios detectado con los datos de *Twitter* y los datos oficiales del número de estudiantes por universidad muestra un ajuste elevado, de 0,95 a nivel de universidad y 0,92 a nivel de campus, lo que indica que la distribución espacial es muy próxima a la realidad. La Figura 38 muestra la distribución de la población universitaria según los datos oficiales y según la muestra de usuarios detectados en *Twitter*. De estos 5.300 usuarios se encontró el lugar de residencia de 2.912 usuarios (el 55% de los posibles usuarios de campus universitarios encontrados). De nuevo, el coeficiente de correlación entre los usuarios de *Twitter* con residencia estimada y los datos oficiales de estudiantes por universidades muestra un ajuste elevado (0,96 a nivel de universidad y 0,91 a nivel de campus).

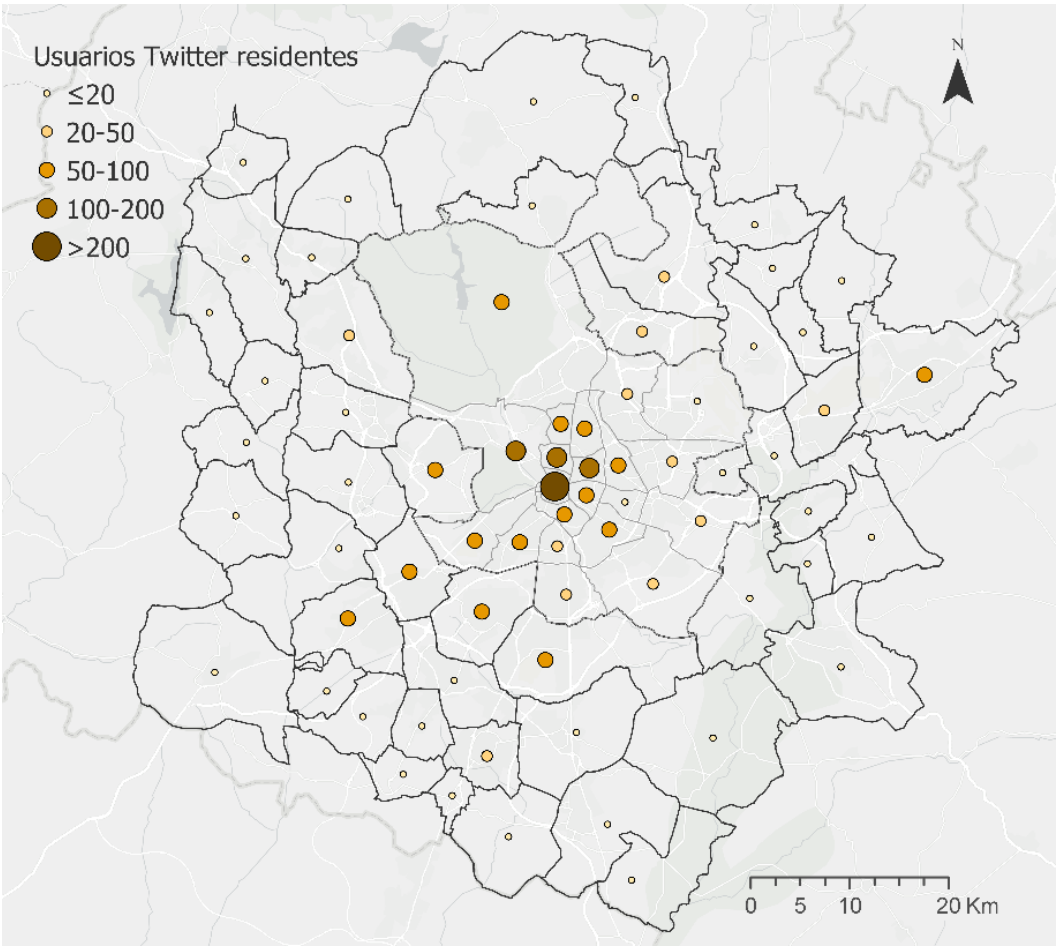
Figura 38: Distribución por campus de los alumnos universitarios a partir de datos oficiales (izquierda) y número de usuarios detectados en *Twitter* (derecha).



Fuente: Elaboración propia a partir de datos del Ministerio de Educación, Cultura y Deporte, y *Twitter*.

Los principales lugares de procedencia de los usuarios identificados con *Twitter* son los distritos centrales de la Almendra Central de Madrid, en especial los distritos de Centro y Chamberí (distritos con barrios y servicios orientados a la población joven). También destaca el distrito de Moncloa-Aravaca (distrito donde se halla el campus de Ciudad Universitaria, y que cuenta con un importante número de residencias universitarias). Además, hay un número importante de residentes en distritos periféricos del norte y sur de la ciudad (zonas dormitorio de la ciudad que albergan campus universitarios como Fuencarral-El Pardo y Vallecas, o se encuentran cerca como Carabanchel). Finalmente, hay que mencionar otros municipios con un elevado número de habitantes y que cuentan con campus universitarios, como Alcalá de Henares, o los principales municipios del sur del área metropolitana (Getafe, Leganés, Alcorcón, Móstoles, y Fuenlabrada) (Figura 39).

Figura 39: Número de usuarios residentes detectados en el Área Metropolitana de Madrid a partir de *Twitter*.

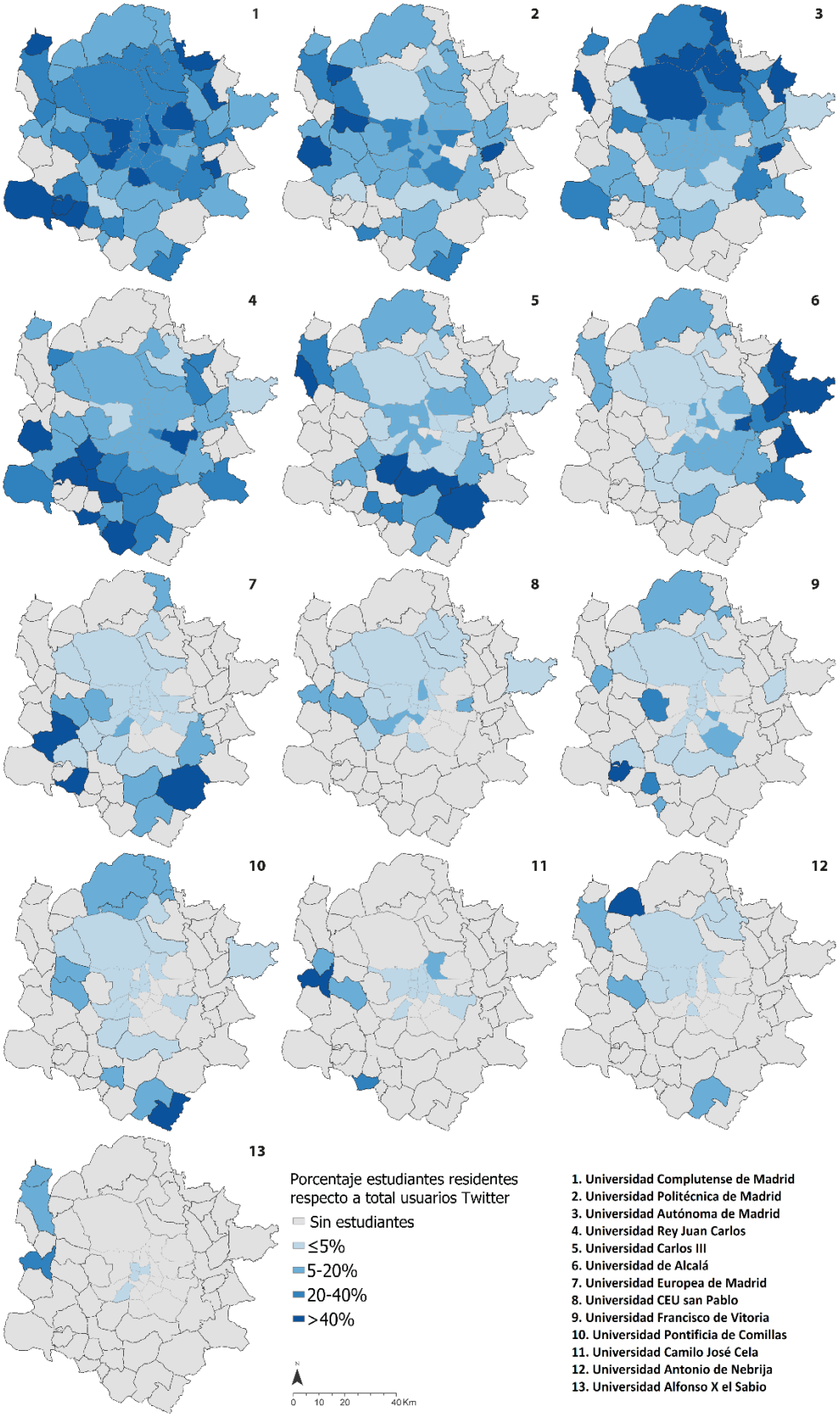


Fuente: Elaboración propia a partir de datos de *Twitter*.

Es posible cartografiar las áreas de influencia de las distintas universidades a partir de los datos de usuarios residentes en cada distrito o municipio. La Universidad Complutense es la que cuenta con una mayor área de influencia que incluye casi toda el área metropolitana. El área de influencia de la Universidad Politécnica es similar en cuanto a extensión, pero cuenta con porcentajes menores de usuarios. La Universidad Autónoma cuenta con una influencia importante en el norte del área metropolitana, mientras que las universidades Rey Juan Carlos y Carlos III destacan en el sur metropolitano y la Universidad de Alcalá al este en el corredor del Henares. Las universidades privadas tienen áreas de influencia menores y con porcentajes de alumnos más bajos respecto a las universidades públicas. Estas áreas de influencia se ubican principalmente en el oeste metropolitano (Figura 40).

En prácticamente todos los casos, destaca la importancia de la proximidad de la residencia respecto a la universidad, coincidiendo a grandes rasgos la ubicación del campus con el área de influencia de la universidad: el campus principal de la Universidad Autónoma está situado en el norte del municipio de Madrid, los campus de las universidades Rey Juan Carlos y Carlos III se hayan en los municipios poblados del sur metropolitano, la Universidad de Alcalá se sitúa en el Corredor del Henares al este, y los principales campus periféricos de las universidades privadas están en el oeste metropolitano.

Figura 40: Porcentaje de población universitaria detectada en *Twitter* por universidad en cada municipio y distrito.



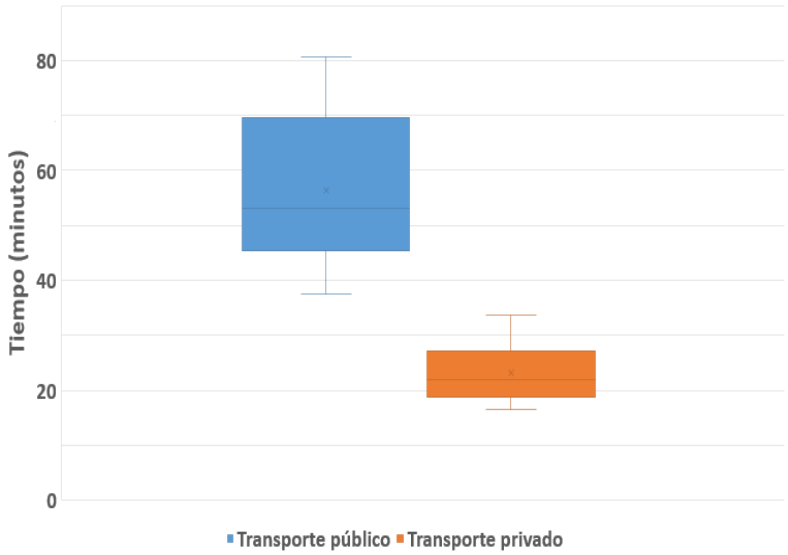
Fuente: Elaboración propia a partir de datos de *Twitter*.

4.3.4. Comparación de tiempos de viajes a partir de transporte público o privado

Los tiempos ponderados obtenidos a partir de datos de *TomTom* para el uso del coche y de ficheros *GTFS* para el sistema de transporte público muestran que hay una gran diferencia en los tiempos de viaje en transporte privado respecto al transporte público: 23,12 minutos de media en coche, mientras el tiempo mediante transporte público aumenta a 56,34 minutos. Las universidades con campus ubicados en el centro, principalmente las universidades Complutense y Politécnica, están mejor conectadas tanto en transporte público como por carretera respecto a las universidades periféricas, por lo que cuentan con tiempos menores de acceso. Mientras, las universidades cuyos campus se sitúan en zonas periféricas del área de estudio (Alcalá, Alfonso X el Sabio), muestran mayores tiempos de viaje (Tabla 18).

Se puede apreciar mayor dispersión de tiempos de viaje a partir de transporte público, mientras que los tiempos de viaje por transporte privado están muy concentrados. Esto se debe a que mientras que el acceso a la red de carreteras del Área Metropolitana de Madrid es homogéneo, hay una diversa variedad de situaciones a la hora de viajar en transporte público a diferentes universidades (variedad de tipos de transporte público, diversidad de horarios y flotas de vehículos, proximidad al tejido urbano, etc.) (Figura 41).

Figura 41: Diagrama de caja de tiempos ponderados de viaje a universidad por tipo de transporte.



Fuente: Elaboración propia a partir de ficheros de transporte *GTFS* y datos de *TomTom*.

Tabla 18: Tiempo medio ponderado a universidades por tipo de transporte.

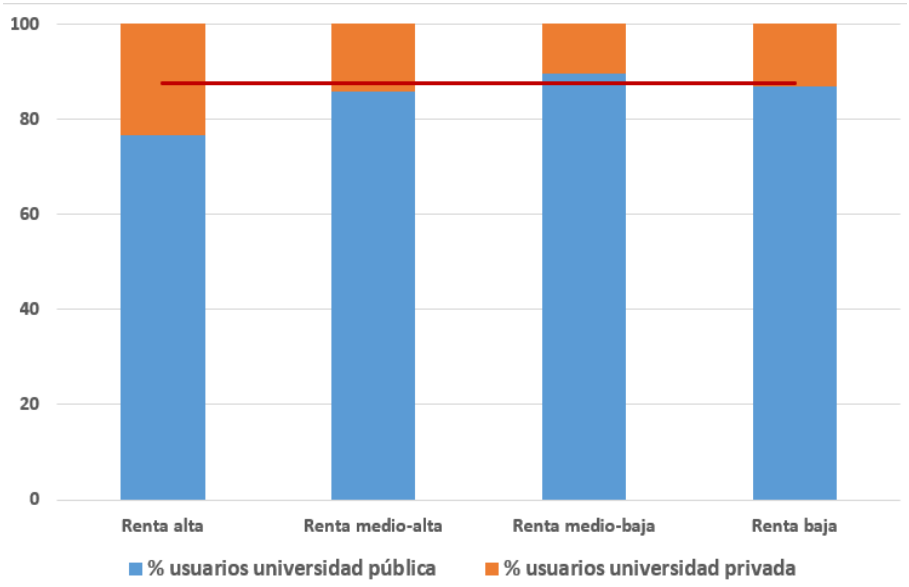
Universidad	**Tiempo medio ponderado en transporte público (minutos)**	**Tiempo medio ponderado en transporte privado (minutos)**
Universidad Complutense de Madrid	38,24	16,61
Universidad Politécnica de Madrid	37,47	16,94
Universidad Autónoma de Madrid	53,06	26,26
Universidad Rey Juan Carlos	58,81	22,65
Universidad Carlos III	48,93	19,62
Universidad Alcalá	80,58	33,76
Universidad Europea de Madrid	70,77	23,92
Universidad CEU San Pablo	45,97	18,25
Universidad Francisco de Vitoria	68,43	21,48
Universidad Pontificia de Comillas	49,04	21,85
Universidad Camilo José Cela	57,31	27,86
Universidad Antonio de Nebrija	45,05	19,18
Universidad Alfonso X el Sabio	78,82	32,20
Total	56,34	23,12

Fuente: Elaboración propia a partir de ficheros de transporte *GTFS* y datos de *TomTom*.

El porcentaje de usuarios de *Twitter* asignados a una universidad pública es del 87,5% mientras que solo el 12,2% de los usuarios encontrados asisten a una universidad privada. Analizando los distritos y municipios del área de estudio clasificados por cuartiles según su nivel de renta, se puede observar que el grupo formado por los municipios de mayor nivel de renta cuenta con el mayor porcentaje de usuarios en universidades privadas (un 23,3%). En el resto de cuartiles se puede observar que los municipios con un nivel de media medio-alto también cuenta con un porcentaje de alumnos asistentes a universidades públicas ligeramente por debajo de la media, mientras que los municipios con nivel de renta medio-bajo o bajo cuentan con un porcentaje de alumnos de universidades públicas superior a la media (Figura 42).

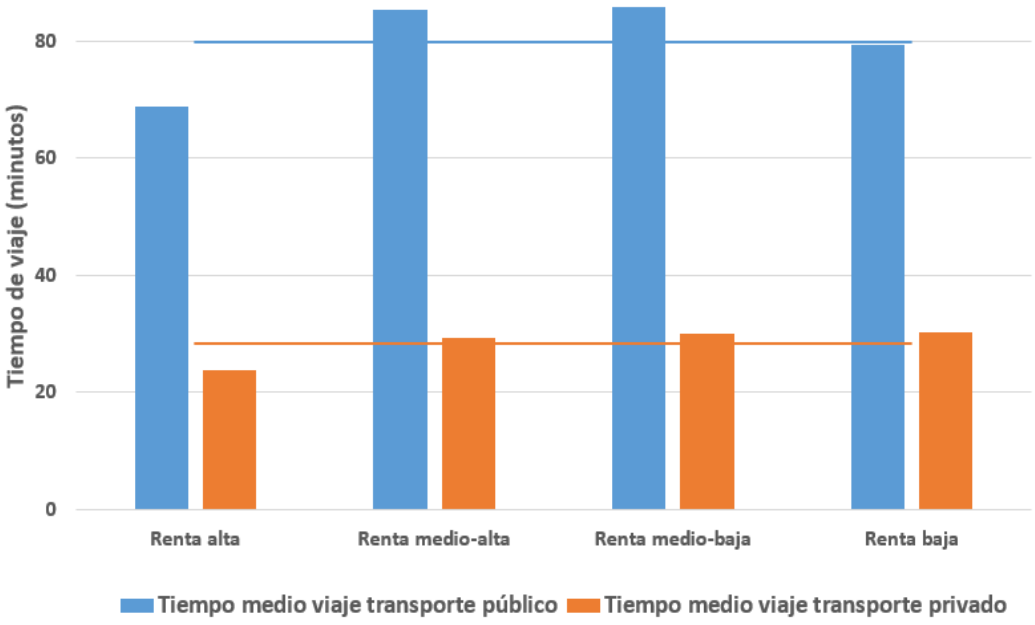
Al analizar los tiempos medios de viaje a partir de los cuartiles de municipios y distritos clasificados según nivel de renta se puede apreciar que aquellos lugares con mayor nivel de renta presentan los tiempos de desplazamiento más bajos. Tanto los tiempos en transporte privado como público del cuartil correspondiente a los municipios y distritos con un nivel alto de renta están por debajo de la media. En el resto de cuartiles se observa que los tiempos de viaje a partir de transporte público son superiores a la media, mientras que los tiempos de viaje usando transporte privado están ligeramente sobre la media (Figura 43).

Figura 42: Porcentaje de alumnos residentes por tipo de universidad según grupos de municipios y distritos por nivel de renta.



Fuente: Elaboración propia a partir de datos de *Twitter*.

Figura 43: Tiempos de transporte privado y público según cuartiles de municipios y distritos por nivel de renta.



Fuente: Elaboración propia a partir de datos de *Twitter*.

4.3.5. Modelo de Huff y relación con la asignación de usuarios de Twitter

El modelo gravitacional de *Huff* asume que los usuarios tienden a escoger la localización de un servicio, en este caso una universidad, basándose en la distancia a ese destino y su capacidad de atracción (representado en este caso por el número de plazas de una universidad). Este modelo estima una probabilidad de usar cada uno de los destinos ofertados (Rodrigue, Comtois, & Slack, 2016). Obviamente, en la elección de la universidad en la que se estudia participan otros muchos motivos. En este trabajo, comparando los modelos resultados de la asignación de usuarios a los campus según *Twitter* y los asignados por el modelo de *Huff*, se quiere ver qué universidades están atrayendo más población universitaria y cuáles menos de los estimados por *Huff*, es decir, la influencia de esos otros motivos.

Para ello se ha calculado el número de usuarios por municipio y distrito y universidad a partir de este modelo gravitacional. Se han obtenido modelos basados en los dos tipos de transporte tratados con unos porcentajes de asignación muy similares, aunque con algunas diferencias. Se puede observar un número de universidades cuyo porcentaje de personas

asignadas en transporte público es mayor respecto al transporte privado (Politécnica, Rey Juan Carlos, Alcalá, Europea de Madrid, Francisco de Vitoria), y universidades con mayores porcentajes asignados en transporte privado respecto al transporte público (Complutense, Autónoma, Carlos III, la mayoría de las universidades privadas) (Tabla 19). El número de plazas ofertadas, la ubicación de la universidad y la proximidad a redes de carreteras o de servicios de transporte público influyen en estos porcentajes.

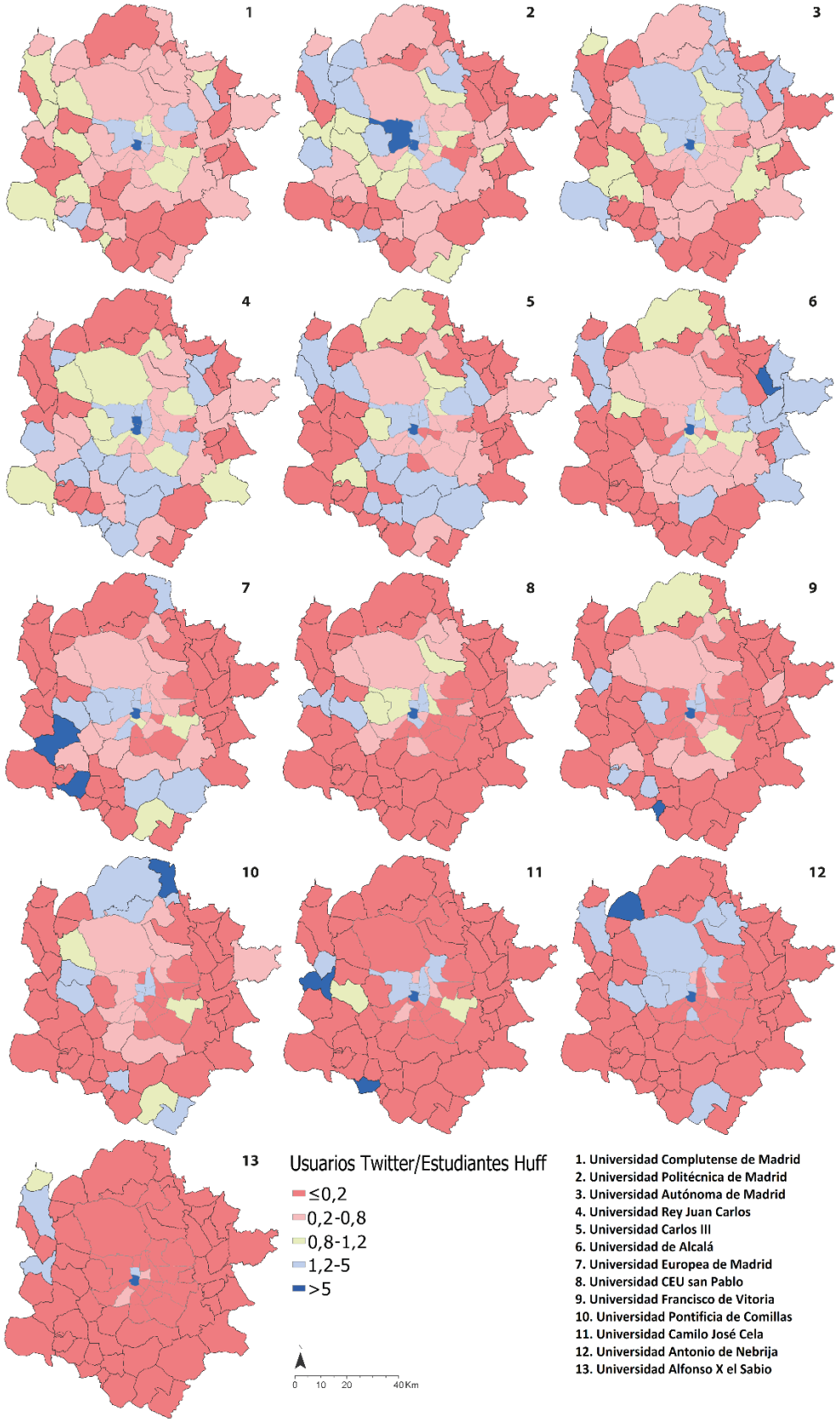
A través de la relación entre usuarios de *Twitter* y la población universitaria asignada al modelo de *Huff* por universidad desde cada municipio y distrito (Figura 44), se observa que en todos los casos en los distritos centrales de Madrid el número de usuarios de *Twitter* es superior al número asignado por el modelo de *Huff*. Igualmente, en cada universidad se observa un número mayor de usuarios de *Twitter* respecto al modelo gravitacional en los municipios y distritos con un campus perteneciente a la universidad de estudio, y en los municipios y distritos adyacentes. En cambio, se puede apreciar un número mayor de usuarios asignados al modelo de *Huff* respecto a *Twitter* en los municipios periféricos del área de estudio. Finalmente, se observa una mayor aproximación entre usuarios de *Twitter* y usuarios asignados al modelo de *Huff* en las universidades públicas, mientras que las universidades privadas presentan en su mayoría un número bajo de usuarios de *Twitter* respecto al número obtenido a través del modelo de *Huff*.

Tabla 19: Porcentaje de estudiantes asignados por universidad a partir del modelo de *Huff*.

Universidad	**% transporte público**	**% transporte privado**
Universidad Complutense de Madrid	27,08	27,92
Universidad Politécnica de Madrid	15,79	15,21
Universidad Rey Juan Carlos	16,18	14,69
Universidad Autónoma de Madrid	9,36	10,66
Universidad Carlos III	6,58	7,02
Universidad de Alcalá	4,45	4,06
Universidad Europea de Madrid	4,46	3,85
Universidad Camilo José Cela	3,17	3,83
Universidad CEU San Pablo	3,60	3,57
Universidad Pontificia de Comillas	2,68	2,93
Universidad Alfonso X el Sabio	2,43	2,53
Universidad Francisco de Vitoria	2,85	2,26
Universidad Antonio de Nebrija	1,38	1,48

Fuente: Elaboración propia a partir de datos de *Twitter*.

Figura 44: Relación entre número de usuarios *Twitter* y estudiantes estimados a partir del modelo de *Huff*.



Fuente: Elaboración propia a partir de datos de *Twitter*.

4.4. Análisis de eventos mediante datos de *Twitter*. El caso de la *World Pride 2017*

4.4.1. Los megaeventos y el Big Data

Los eventos de masas son fenómenos con alta repercusión y seguimiento por parte de la población debido a su magnitud, importancia en la sociedad y publicidad. Los megaeventos se pueden definir como eventos a gran escala que están orientados especialmente al mercado del turismo internacional. Destacan por su tamaño en términos de asistencia, financiación pública, cobertura televisiva, construcción de infraestructuras e impacto en los residentes de la ciudad o país organizador (Knott et al., 2015). Eventos a gran escala como los Juegos Olímpicos, exposiciones mundiales, conciertos, etc. ocurren cada año en todo el mundo, atrayendo a un gran número de participantes y turistas que viajan a un destino concreto, principalmente una gran ciudad (Xu & González, 2017).

Los festivales y grandes eventos tienen una gran importancia para la administración y organización de una ciudad. El principal motivo es el impacto económico que generan los turistas a partir del consumo que realizan en la ciudad (Batista e Silva et al., 2018). Los efectos de la celebración de estos eventos se dan en varias áreas (económica, comercial, infraestructuras, política, medioambiental, etc.). Estos efectos pueden ser positivos (aumento del turismo, de las oportunidades de empleo, reconocimiento y visibilidad a nivel internacional, atracción de capital y empresas, desarrollo de infraestructuras, revitalización de áreas urbanas, aumento de la seguridad, oportunidades de participación, atención al medioambiente, etc.), pero también negativos (centralización de recursos, aumento de los precios, aumento de la especulación inmobiliaria, generación de desechos, problemas de congestión y contaminación, pérdida de control de recursos en detrimento de las empresas privadas, contratos basura, etc.) (Knott et al., 2015).

Las empresas y administraciones públicas necesitan de datos e información que permitan realizar diagnósticos eficaces del impacto que tiene un evento sobre la ciudad, y así poder facilitar su organización y gestión. Como ya se ha comentado previamente en esta tesis, adquirir una alta cantidad de datos en un área geográfica usando métodos convencionales consume tiempo y es caro (N. Caceres et al., 2007). La participación de los ciudadanos como generadores de datos se antoja valiosa para los organismos públicos y privados, ya que la huella digital generada puede ayudar al diseño y funcionalidad de la ciudad y a la administración de eventos (Pallares-Barbera & Masala, 2016). Un ejemplo del uso de fuentes de datos masivos para la administración y gestión de eventos se halla en el uso

del *Big Data* por parte del Ayuntamiento de Sevilla para contabilizar el número de desplazamientos a pie en la ciudad durante la Semana Santa[18].

Los eventos, sobre todo los más importantes a nivel mundial (pero también los festivales de menor entidad), cuentan con un seguimiento que se magnifica en las redes sociales. En España, durante el año 2016, las dos principales búsquedas realizadas en *Google* fueron sobre los Juegos Olímpicos de Río de Janeiro y la Eurocopa celebrada en Francia. Mientras, en *Twitter*, el *trending topic* o tema más comentado a lo largo del año ha sido el Festival de Eurovisión celebrado en Kiev, seguido de los Juegos Olímpicos y la Eurocopa en cuarto lugar. Además, las organizaciones usan las redes sociales como un medio fundamental para publicitar los eventos de masas con el fin de conseguir una mayor repercusión (Leung, Law, van Hoof, & Buhalis, 2013).

Se pueden observar dos tipos de eventos a partir de sus características en redes sociales como *Twitter*. Una primera categoría englobaría a los eventos globales, destacando su gran volumen de mensajes, y una mayor importancia del tema del evento respecto a la ubicación (un ejemplo sería la *World Pride*, evento a nivel mundial cuya importancia radica en su temática). El segundo tipo de eventos se sitúa a nivel local, con una cantidad menor de mensajes, y donde la ubicación es tan importante como la temática (los festivales de música destacan tanto en su temática como en el hecho de que es una festividad centrada en la propia ciudad) (Liu, Ge, Zheng, Lin, & Li, 2018).

El objetivo principal de este caso de estudio es analizar el impacto que ha tenido la *World Pride* 2017 en la ciudad de Madrid mediante la identificación de la procedencia de los visitantes, y de la comparación del fenómeno con una semana habitual a partir de datos geolocalizados de *Twitter*. Usando técnicas de geoestadística y geovisualización se pueden cartografiar y comparar los resultados obtenidos tanto en el evento como en la semana habitual, y evaluar el impacto espacio-temporal del festival, partiendo de la hipótesis de que cuanto mayor es la densidad de la gente en un sitio como una plaza durante un concierto, mayor es la posibilidad de usar una herramienta de comunicación como las redes sociales para compartir la actividad, opinión, o posición de una persona (Pallares-Barbera & Masala, 2016).

[18]https://www.sevilla.org/ayuntamiento/alcaldia/comunicacion/noticias/el-estudio-smart-city-sevilla-contabiliza-medio-millon-de-desplazamientos-a-pie-al-dia-en-el-entorno-de-la-carrera-oficial-durante-el-arranque-de-la-semana-santa-2018

4.4.2. Metodología específica para el análisis del impacto de un evento en la ciudad mediante datos de Twitter

La base de datos de *Twitter* geolocalizados empleada en este caso contiene *tweets* ubicados en la Almendra Central de Madrid en el periodo comprendido entre el lunes 26 de junio y el domingo 2 de julio del año 2017. Dicha base de datos inicial consta de 48.175 *tweets* de 14.353 usuarios. Al estar tratando con población turística no habitual o residente en Madrid, solo se aplicó el filtro de eliminación de *bots* respecto a la limpieza de datos de la metodología general. En el periodo de la *World Pride,* en la Almendra Central, se publicaron 18.439 *tweets,* compartidos por 7.406 usuarios. Para comparar los resultados de la *World Pride* con la actividad habitual de una semana próxima, se han tomado los *tweets* escritos en la semana del 12 al 18 de junio, donde se descargó un total de 8.896 mensajes publicados por 4.053 usuarios.

Además, se han identificado los 7.406 usuarios que se detectaron en la *World Pride* para obtener los 3.200 últimos *tweets* de cada uno de ellos, e intentar obtener así más mensajes que no se hubiesen captado en *streaming.* Esta búsqueda permitió también crear una nueva base de datos de *tweets* a escala global donde identificar sus procedencias. Como resultado, se ha obtenido una tercera base de datos formada por 1.215.994 *tweets* con información enriquecida del lugar desde donde fueron publicados a partir de la unión espacial en SIG con capas de países y de provincias de España. Al haber enriquecido la base de datos con el nombre del país desde donde fue publicado cada *tweet*, se pudo extraer el número de *tweets* que ha escrito cada usuario desde cada país donde ha publicado un mensaje, y se estableció como país de origen aquel con un mayor número de mensajes. Para los usuarios cuyo país de origen era España se repitió el proceso a nivel provincial, con el fin de conocer la provincia de origen.

El análisis temporal de los datos se ha efectuado sobre la distribución diaria del número de *tweets* y usuarios agregados a partir de sus identificadores durante el periodo de estudio. Para la identificación del idioma principal del usuario, se ha seleccionado el idioma que aparece en un mayor número de mensajes por usuario. La impronta espacial del evento se analizó a nivel de barrios y de secciones censales mediante el conteo de usuarios localizados por cada unidad, y a través del cálculo del incremento del porcentaje respecto al número de usuarios encontrados en el mismo polígono en la base de datos de la semana habitual. Estos resultados se complementaron con un análisis de autocorrelación espacial (Moran I) para evaluar el grado de agrupación de usuarios en el

evento, y con análisis LISA (*Local Indicators of Spatial Association*) para la visualización de concentraciones de valores altos, bajos, y *outliers* espaciales en la ciudad durante la *World Pride* (Anselin, 1995).

La distribución de la población en la ciudad varía a diferentes horas del día y puede ser analizada a partir de la huella digital de la población a cada hora del día en cada punto de la ciudad (García-Palomares et al., 2018). El análisis espacio-temporal de este caso de estudio se ha realizado a nivel de secciones censales, calculando el número de usuarios que ha *twitteado* en cada sección por franjas horarias. La visualización se ha realizado por mapas de símbolos proporcionales a partir de dos animaciones que muestran por distintos momentos del día los cambios en el número de usuarios detectados en el centro de Madrid durante la semana del evento y una semana habitual. Cada franja es de 6 horas, dando como resultado visualizaciones en cuatro momentos del día (mañana, mediodía, tarde, y noche).

Se puede visualizar la impronta semántica del evento comparando palabras relacionadas con la temática del festival durante los dos periodos temporales de estudio a partir del conteo del número de *tweets* que incluyen estas palabras y la sumarización por el campo del identificador del usuario. Como resultado, el porcentaje de *tweets* que incluyen estas palabras temáticas es significativamente mayor durante el evento (Tabla 20).

Tabla 20: Palabras claves relacionadas con la *World Pride* por usuarios de *Twitter*.

Palabra	Numero usuarios *World Pride*	% usuarios *World Pride*	Numero usuarios semana normal	% usuarios semana normal
Pride	1480	19,98	33	0,81
World Pride	695	9,38	14	0,35
Orgullo	679	9,17	11	0,27
Gay	293	3,96	30	0,74
Chueca	272	3,67	28	0,69
LGTB	87	1,17	8	0,20
LGBT	73	0,99	7	0,17

Fuente: Elaboración propia a partir de datos de *Twitter*.

4.4.3. Carácter multicultural del evento. Lugares de procedencia de los visitantes

Una primera aproximación al carácter multicultural del evento se ha realizado mediante la comparación de la distribución de los idiomas principales de los usuarios de las dos muestras de *Twitter*. Mientras en una semana habitual aparecen 26 idiomas entre los *tweets* recogidos en la Almendra Central, durante la *World Pride* esta cifra aumenta hasta 32 idiomas diferentes. El cambio drástico sucede en el barrio de Chueca, donde los 9 idiomas utilizados en la semana habitual contrastan con las 21 lenguas registradas durante la *World Pride* Los idiomas más encontrados han sido el castellano y el inglés, seguidos por el portugués, el francés, y el italiano (Tabla 21).

Tabla 21: Número de usuarios según idioma configurado en *Twitter.*

Idioma	**Almendra Central *World Pride***	**%**	**Almendra Central semana normal**	**%**	**Chueca *World Pride***	**%**	**Chueca semana normal**	**%**
Castellano	5.073	68,50	2.856	70,47	882	55,82	206	71,28
Inglés	1.860	25,11	836	20,63	574	36,33	50	17,30
Portugués	149	2,01	83	2,05	18	1,14	5	1,73
Francés	60	0,81	29	0,72	9	0,57	3	1,04
Italiano	50	0,68	16	0,39	9	0,57	0	0,00
Otros	214	2,89	233	5,75	88	5,57	25	8,65
Totales	7.406	100,00	4.053	100,00	1.580	100,00	289	100,00

Fuente: Elaboración propia a partir de datos de *Twitter*.

La comparación de la distribución de idiomas en la Almendra Central durante la *World Pride* y durante una semana habitual permite apreciar el aumento del porcentaje de las lenguas extranjeras empleadas durante el evento, destacando el idioma inglés (25% de presencia durante el festival frente al 20% durante una semana habitual). Este fenómeno es más visible en el barrio de Chueca donde en la *World Pride* hay un 36% de mensajes escritos en inglés, frente a un 17% en una semana habitual. El caso contrario sucede con el idioma español, con registros muy similares en las dos semanas de estudio en la

Almendra Central (en torno al 70%), pero con un porcentaje ligeramente menor en la *World Pride*. De nuevo, esta situación se visualiza mejor en el barrio de Chueca; durante la *World Pride* (un 56%) el uso del idioma español es menor respecto a su uso en una semana habitual (un 71%).

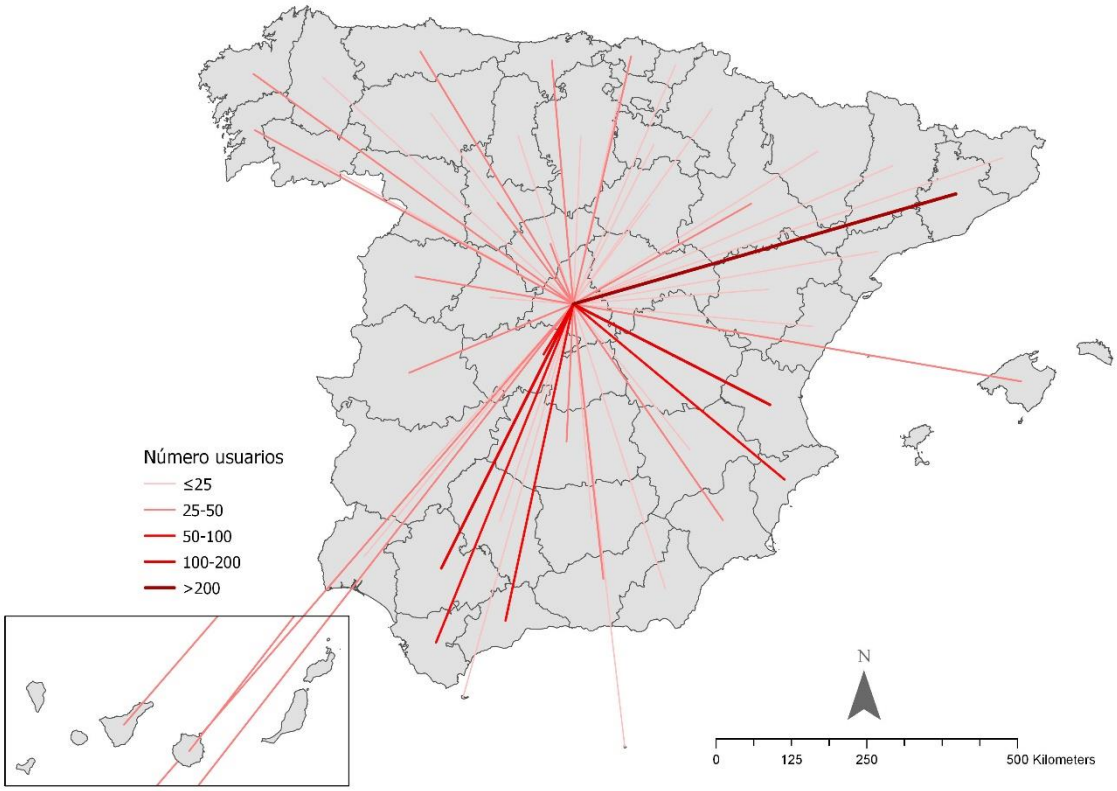
Respecto a la procedencia de los visitantes de la *World Pride*, a nivel nacional destaca la propia Comunidad de Madrid cómo principal lugar de origen, resultado que coincide con la información suministrada por el propio Ayuntamiento de Madrid[19]. En cuanto a otras provincias, en segundo lugar, se sitúa Barcelona, seguida de las provincias más pobladas del litoral mediterráneo (Valencia, Alicante), Andalucía (Sevilla, Cádiz, Málaga), las Islas Canarias, y provincias cercanas cómo Toledo (Figura 45). Se puede resumir que los visitantes españoles han viajado preferentemente desde las áreas metropolitanas más pobladas del país, o desde las ciudades más cercanas a Madrid.

A nivel internacional los principales países de origen son de América, destacando Estados Unidos, y los principales países de América Latina (México, Brasil, Colombia, Perú, y Argentina). La segunda gran zona de procedencia de los visitantes es Europa Occidental, donde destaca Reino Unido, seguido de Italia, Francia, Portugal, y Alemania (Figura 46). Estos resultados coinciden con los obtenidos por el buscador de viajes *GoEuro*. Esta empresa permite comprar o reservar viajes a partir de la selección de un lugar de destino y una fecha de desplazamiento. Según sus datos, los países con mayor número de visitantes durante la *World Pride* fueron también Estados Unidos y los países de América Latina y Europa Occidental. En cuanto a rutas nacionales, la empresa *GoEuro* destaca las conexiones con Barcelona, Valencia, y Sevilla (las provincias con mayor registro de usuarios locales según la Figura 45). En cuanto a los trayectos internacionales, los más importantes fueron con las ciudades de París, Roma, y Lisboa[20].

[19]https://diario.madrid.es/blog/notas-de-prensa/el-gasto-en-el-centro-de-la-ciudad-aumento-un-15-durante-el-world-pride/

[20]https://www.hosteltur.com/comunidad/nota/019268_argentinos-y-colombianos-los-mas-orgullosos-de-viajar-al-worldpride-2017.html

Figura 45: Usuarios detectados en *Twitter* durante la *World Pride* según provincia de procedencia.



Fuente: Elaboración propia a partir de datos de *Twitter*.

Figura 46: Usuarios detectados en *Twitter* durante la *World Pride* según país de procedencia.

Fuente: Elaboración propia a partir de datos de *Twitter*.

4.4.4. La impronta espacio-temporal del evento

Los resultados obtenidos muestran un mayor volumen de mensajes y de usuarios en la semana de la *World Pride* que durante el periodo comprendido por las dos semanas previas o las dos semanas posteriores al evento. Se observa un paulatino crecimiento conforme avanza el festival (Figura 47a). El momento cumbre de este crecimiento se da el fin de semana del 1 y 2 de julio, día de la clausura del evento, con un incremento del 88% de *tweets* publicados y del 71% de usuarios respecto al promedio del periodo analizado. Una vez terminada la *World Pride*, se aprecia un descenso de la actividad en *Twitter*, indicativo de que los visitantes se van de la ciudad tras haber terminado el evento, pero también de que los propios ciudadanos madrileños publicaron un mayor número de mensajes en la fiesta y su actividad en *Twitter* es menor en una semana habitual.

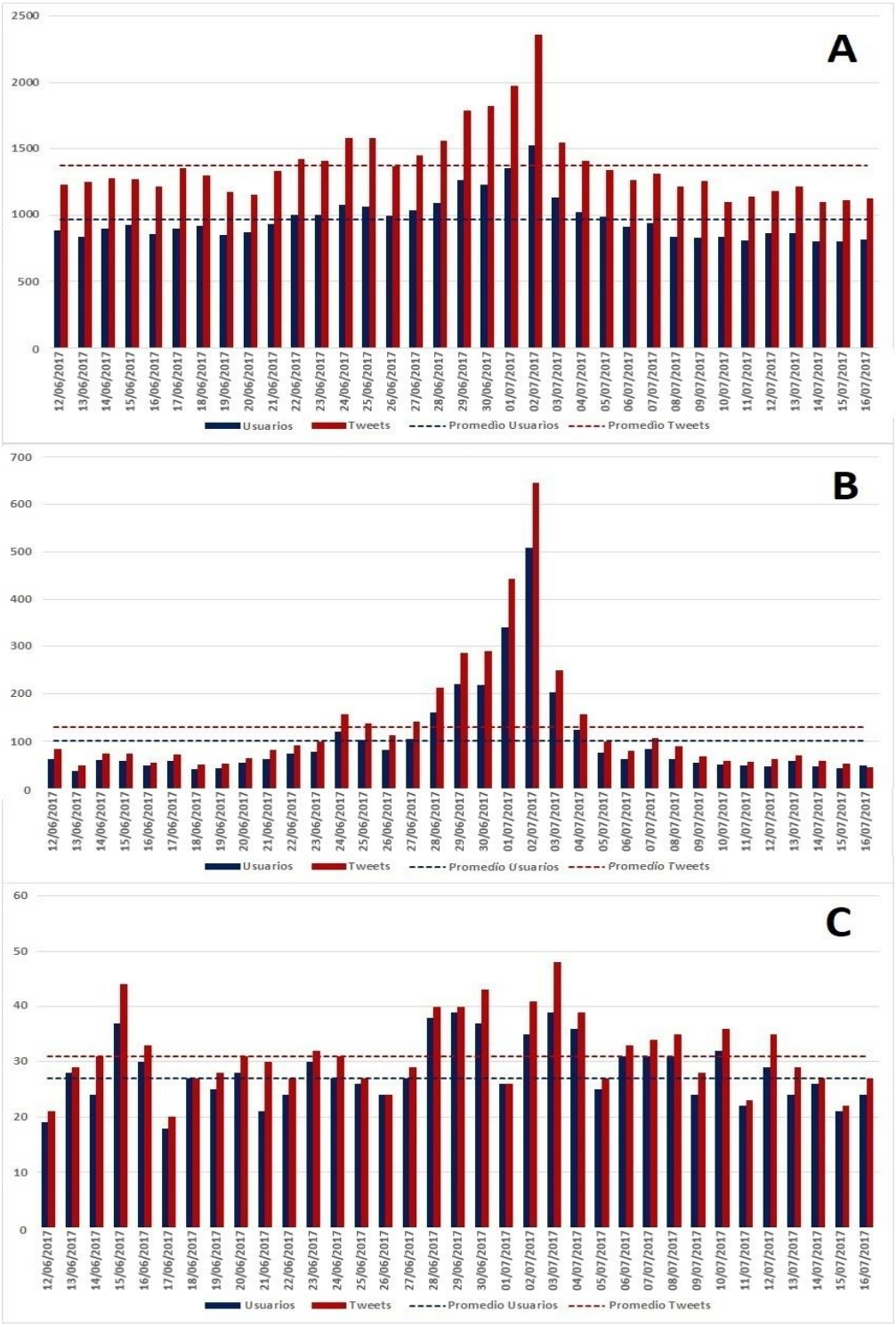
Si se reduce el ámbito espacial al barrio de Chueca, se puede observar con mayor claridad un incremento de la actividad de *Twitter* en la franja correspondiente a las fechas de la *World Pride* (Figura 47b). En este caso hay un aumento brusco de la actividad durante todo el evento en comparación con el bajo volumen de *tweets* en los días anteriores o posteriores. Así, en los días 1 y 2 de julio el incremento respecto al promedio sube al 706% en cuanto a *tweets*, y al 695% de usuarios. Además, el lunes 3 de julio (el día después del final de la *World Pride*) fue el día con mayor número de mensajes enviados en el barrio del Aeropuerto de Barajas (Figura 47c), con un aumento del 77% de *tweets* y del 54% de usuarios respecto a la media.

Analizando la actividad por número de usuarios según la distribución horaria de una semana habitual y comparándola con la actividad registrada en la semana de la *World Pride*, en la Almendra Central se visualiza como la curva de actividad tiene a grandes rasgos el mismo perfil, con la diferencia de que en la *World Pride* hubo mayor número de usuarios (Figura 48a). En ambas situaciones, los momentos de mayor actividad se dieron durante la tarde (de 12 a 16) y la noche (de 20 a 23), situándose el pico a las 22 de la noche. A esta hora, el número de usuarios durante el festival tuvo un 28% de mayor actividad respecto a la semana habitual.

Sin embargo, al analizar la actividad por horas en el barrio de Chueca, las curvas entre una semana habitual y el *World Pride* son muy diferentes. En la recta de la *World Pride* destacan dos grandes picos de actividad: de 13 a 15 de la tarde y de 21 a 23 de la noche, causando una gran diferencia de actividad respecto a una semana habitual donde se ve un

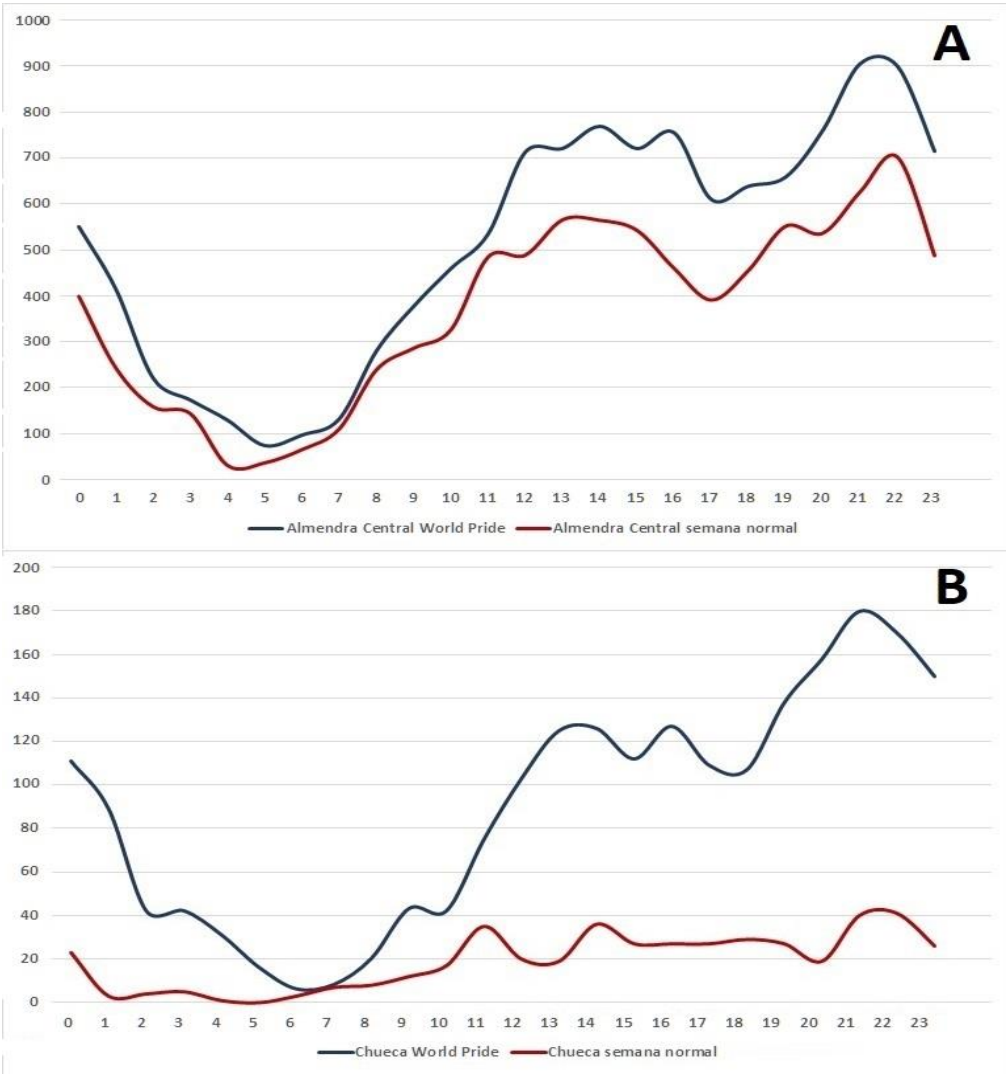
patrón más lineal y con pocos crecimientos. Durante la hora cumbre, las 22 horas, el incremento de actividad fue del 350% de usuarios respecto a la semana habitual (Figura 48b).

Figura 47: Volumen de actividad por días en la Almendra Central (A), Barrio de Chueca (B), y Aeropuerto (C).



Fuente: Elaboración propia a partir de datos de *Twitter*.

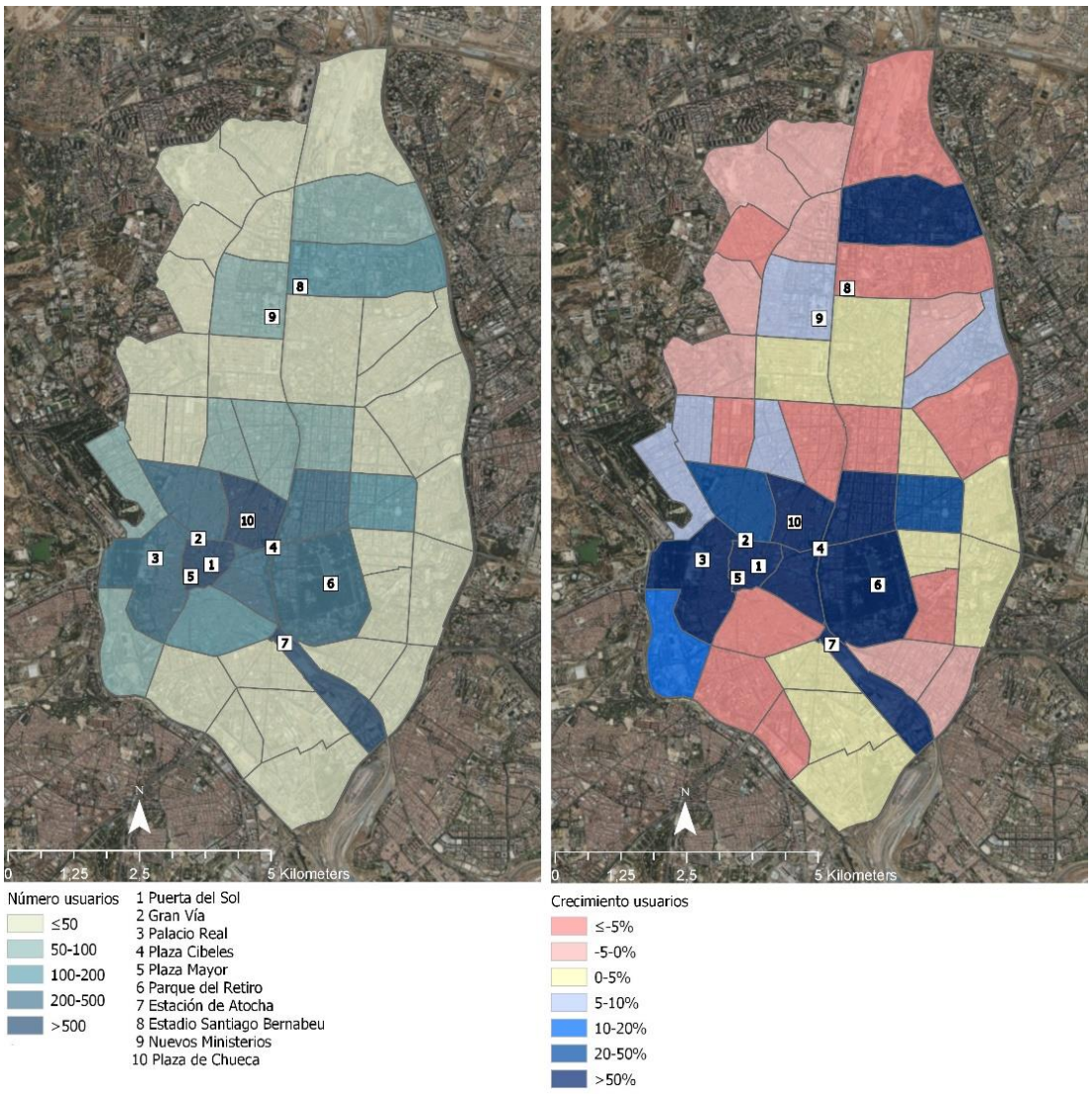
Figura 48: Comparación de número de *tweets* por horas en la Almendra Central (A) y en el Barrio de Chueca (B).



Fuente: Elaboración propia a partir de datos de *Twitter*.

Espacialmente, las zonas con mayor número de usuarios localizados durante la celebración de la *World Pride* coinciden con los principales lugares que acogieron actividades del evento. A nivel de barrios se pueden apreciar como los barrios con un mayor número de usuarios detectados durante el evento fueron Sol (principal centro de la ciudad), Chueca y Atocha. Todo el distrito Centro (excepto el barrio de Lavapiés), y los barrios colindantes (como Retiro) presentan un fuerte aumento del porcentaje de usuarios en comparación con la semana habitual. Igualmente se aprecia como muchos barrios de la Almendra Central que no son colindantes al distrito Centro sufren un porcentaje negativo que indica un decrecimiento de la actividad durante el evento, y una concentración de dicha actividad en el Distrito Centro (Figura 49).

Figura 49: Número usuarios (izquierda) y cambio porcentual (derecha) respecto a semana habitual en la Almendra Central por barrios durante la *World Pride.*

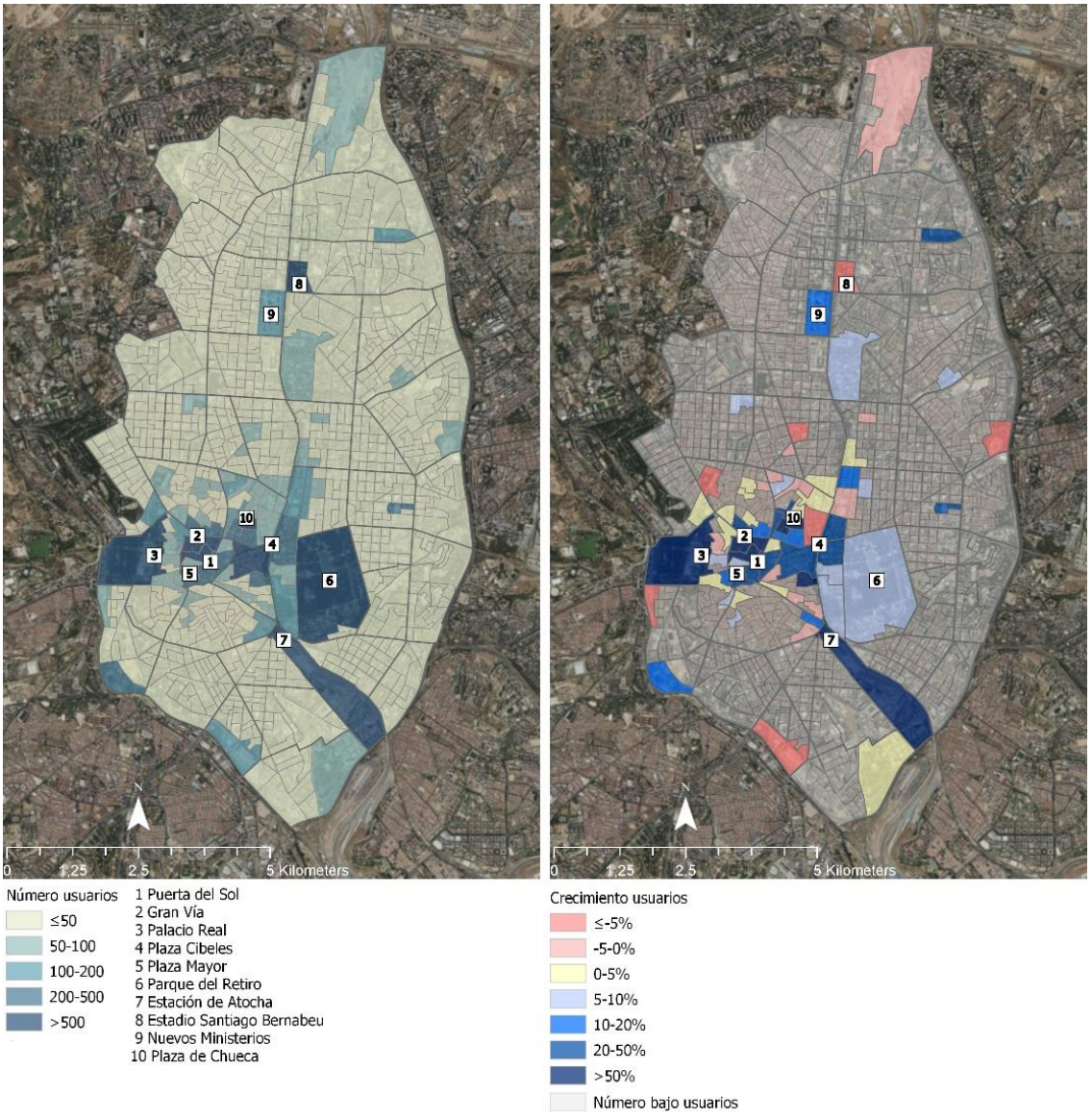


Fuente: Elaboración propia a partir de datos de *Twitter*.

Si se reduce la escala de análisis y se trabaja con secciones censales, se puede ver con mayor detalle como los usuarios se concentran en el eje de la Gran Vía y en las plazas del barrio Chueca. También se observa una importante presencia de usuarios en los principales lugares turísticos del área de estudio (el triángulo Sol-Callao-Opera, Palacio Real, los estadios Santiago Bernabéu y Vicente Calderón, Palacio de los Deportes de la Comunidad de Madrid, Paseo de Recoletos y Parque del Retiro). En cuanto al porcentaje de crecimiento respecto a la semana habitual, se ha optado por ignorar las secciones con menos de 10 usuarios. De esta forma, se visualiza con mayor claridad el fuerte crecimiento de usuarios en el barrio de Chueca y en otros puntos centrales como la Puerta de Sol, Gran Vía, o Palacio Real (los principales puntos turísticos del casco histórico),

además del Palacio de los Deportes de la Comunidad de Madrid, la estación intermodal de Nuevos Ministerios, o la estación de trenes de Atocha. A su vez, hay un porcentaje negativo de crecimiento de usuarios en el estadio Santiago Bernabéu (pese a ser uno de los lugares con mayor número de usuarios), y en otras zonas de especial relevancia como la plaza de toros de Las Ventas, puntos de Madrid Río, el Palacio de Conde-Duque, o la estación de trenes de Chamartín. Por otra parte, el parque del Retiro presenta un crecimiento débil de usuarios respecto a una semana habitual. Esto se debe a que mientras en una semana habitual el parque es un punto de atracción de visitantes, durante el evento, aunque sigue atrayendo usuarios, pierde interés en detrimento de otras áreas (Figura 50).

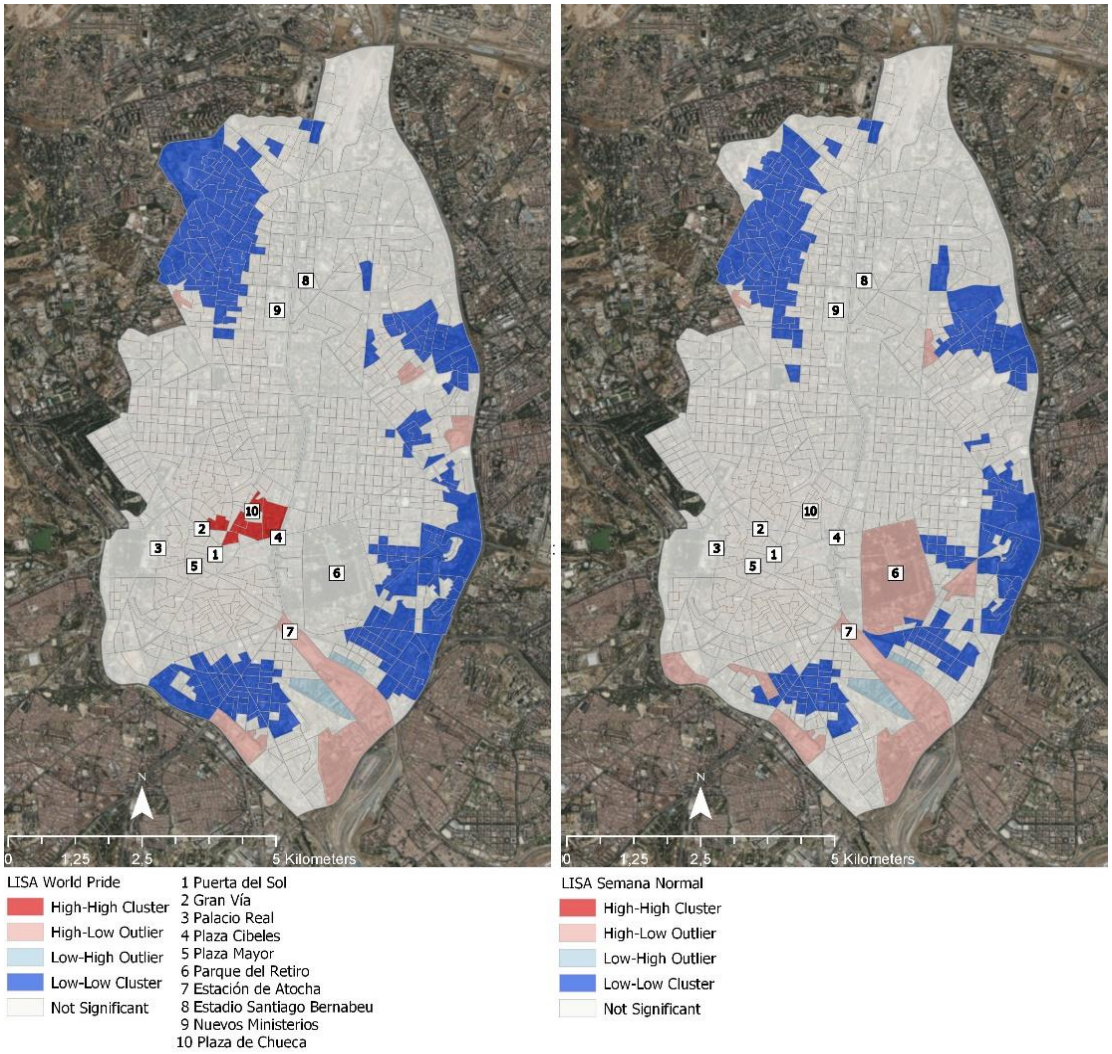
Figura 50: Número usuarios (izquierda) y cambio porcentual (derecha) respecto a semana habitual en la Almendra Central por secciones censales durante la *World Pride*.



Fuente: Elaboración propia a partir de datos de *Twitter*.

El Índice de Moran obtenido durante la *World Pride* (valor z 3,37 y valor p 0,0007) indica la formación de clústeres debido a la fuerte presencia de visitantes en la Almendra Central. Aunque hay también presencia turística durante la semana habitual de estudio, hay una menor clusterización de usuarios respecto a la semana del evento (valor z 2,2 y valor p 0,027). El Índice LISA cartografía la distribución de estos clústeres. Tanto en la semana del festival como en la semana habitual hay un clúster de valor alto rodeado de áreas de valor bajo en la estación de Atocha, y clústeres de poca actividad en el distrito de Tetuán y en barrios situados al este y sur de la Almendra Central. Sin embargo, durante la *World Pride* se genera un clúster de valor alto rodeado de áreas de mucha actividad en el barrio de Chueca, mientras que el valor alto de los clústeres correspondientes al Parque del Retiro, y a Madrid Río en la semana normal desaparecen (Figura 51).

Figura 51: Clústeres de asociación espacial local en la Almendra Central por secciones censales durante la *World Pride* (izquierda) y durante una semana habitual (derecha).

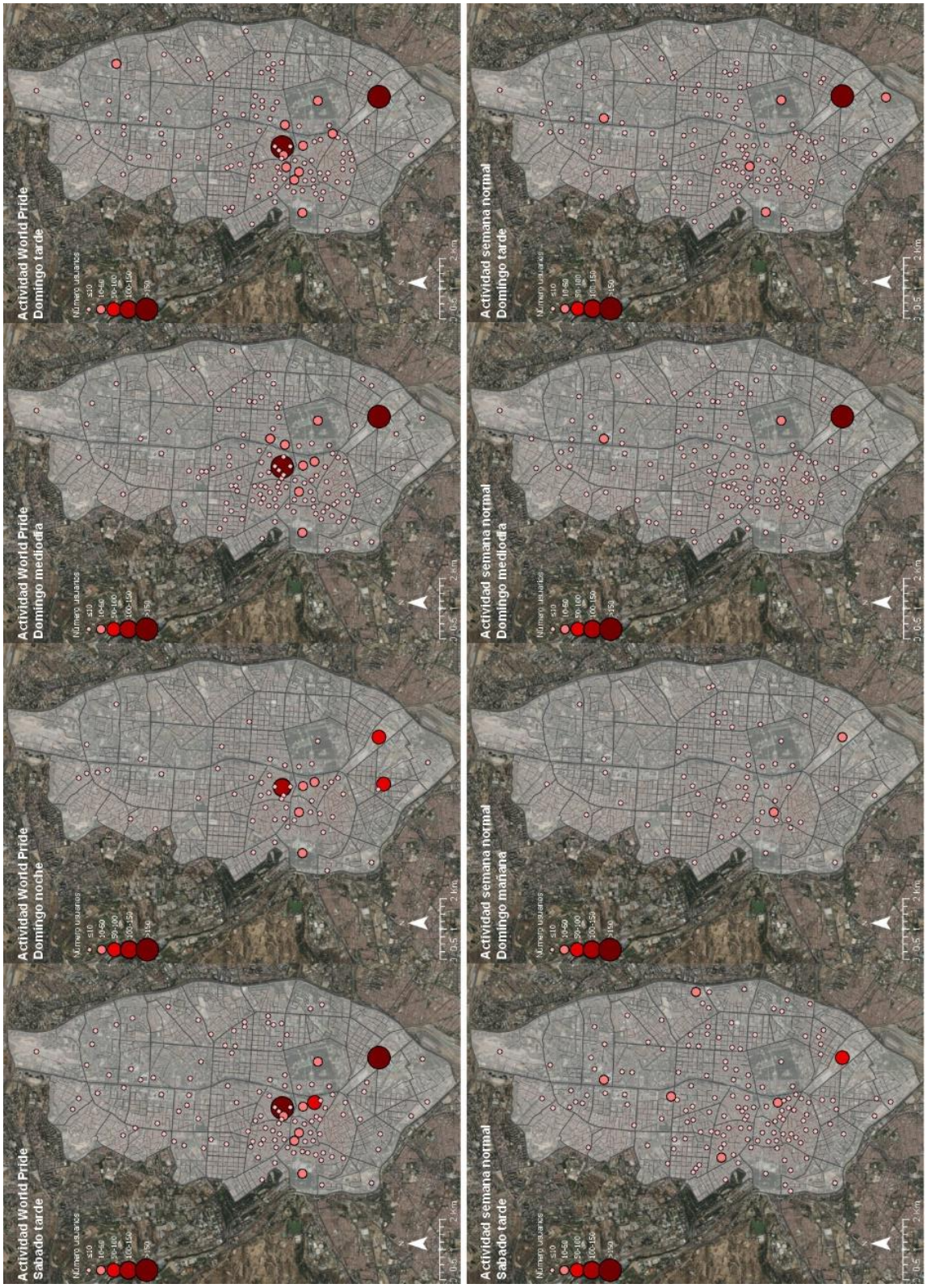


Fuente: Elaboración propia a partir de datos de *Twitter*.

La actividad detectada en ambas casuísticas a través de los datos de *Twitter* varía dependiendo de la hora, el día y el lugar. Estudiando la actividad detectada durante una semana habitual, se observa como el mayor foco de usuarios corresponde con la estación de trenes de Atocha, y luego hay una actividad centrada de manera uniforme en el casco histórico, el parque del Retiro, y en secciones de los distritos de Salamanca y Chamberí. En cambio, se puede apreciar una mayor actividad en estas mismas zonas (un mayor número de círculos, de mayor tamaño, y en un mayor número de franjas horarias) durante la *World Pride*.

Además, mientras que la estación de Atocha es el lugar con mayor número de usuarios durante cualquier momento de la semana habitual, en determinados momentos de la *World Pride* (como la tarde-noche del sábado o el mediodía del domingo) se visualiza un mayor número de usuarios en el barrio de Chueca que en la estación de Atocha. Al contrario, el estadio Santiago Bernabéu es el único lugar con círculos de mayor tamaño durante una semana habitual. El Archivo Multimedia 1 compara de forma animada el número de usuarios detectados en cada franja horaria a lo largo de la semana de la *World Pride* con la actividad detectada en una semana habitual. La Figura 52 incluye un mapa donde se pueden comparar los momentos más importantes del evento (las franjas horarias en fin de semana) con la actividad en esas mismas franjas en una semana normal.

Figura 52: Número de usuarios por secciones censales en la Almendra Central en franjas del fin de semana.



Fuente: Elaboración propia a partir de datos de *Twitter*.

4.5. Evaluación de las percepciones del Metro de Madrid a través de los textos de *Twitter*

4.5.1. La percepción de los sistemas de transporte público

Como se comentó en la introducción de la tesis doctoral, ofrecer un servicio de movilidad equilibrado y sostenible es uno de los mayores desafíos de las áreas metropolitanas. Los sistemas de transporte público ahorran energía y reducen las emisiones contaminantes, por lo que son uno de los principales ejes en el uso de la energía limpia. Promover el transporte público es vital para las ciudades, que necesitan descongestionar el tráfico y reducir el nivel de polución (Hosseini et al., 2018). En España, casi 5 millones de ciudadanos han viajado en transporte público en 2018, representando el 10,48% de la población total del país (el 48,56% de la población si tenemos en cuenta solo los ciudadanos residentes en municipios con más de 300.000 habitantes, que reúnen una cuarta parte de la población del país)[21].

La accesibilidad es un concepto clave en el desarrollo urbano por la capacidad para vincular las actividades de los ciudadanos (usos del suelo) con la movilidad necesaria para realizar estas actividades (usos del transporte) (Moya-Gómez et al., 2017). Uno de los objetivos de la planificación urbana es dar soluciones a los problemas de movilidad de las ciudades mediante la construcción y mejora de las infraestructuras de transportes. Ya se ha mencionado en la introducción de la tesis que el aumento de la demanda de la movilidad ha provocado también un aumento de la congestión en sistemas de transporte público, y la necesidad diseñar un modelo de transporte sostenible que facilite el uso inteligente del transporte público (Banister, 2008). Hace falta diseñar modelos para estimar la demanda del transporte en diferentes escenarios y estudiar la relación entre el sistema de transporte y el sistema territorial (Gutiérrez-Puebla et al., 2019).

En este paradigma, un problema de creciente interés para las agencias que apoyan los servicios de transporte público es disponer de información sobre la funcionalidad de los servicios que ofrecen, para detectar rápidamente problemas como frecuencias y horarios de vehículos, o la necesidad de instalar una parada de Metro en un barrio (Ji et al., 2018). La opinión de los ciudadanos sobre el transporte público es fundamental tanto para los organismos públicos como para las compañías de transporte. Esa información es clave

[21] https://www.statista.com/statistics/772026/number-of-public-transport-passengers-in-spain/

para comprender las necesidades, motivaciones y sensibilidades de los usuarios de transporte público, y para proveer de una visión útil a los futuros modelos de planeamiento urbano que respondan a esas necesidades (El-Diraby et al., 2019).

El *Big Data* es una fuente de datos valiosa para obtener información sobre el funcionamiento de los diferentes modos de transporte. La información voluntaria creada a partir de las aplicaciones de las redes sociales crea un rango de oportunidades para las agencias de transporte público, ya que les permite comprender mejor las necesidades y opiniones de los usuarios que usan sus servicios (Casas & Delmelle, 2017). Para ello, estas agencias adoptan aproximaciones a las redes sociales para conectar y comunicarse con sus usuarios, proporcionándoles de información sobre sus servicios (Manetti, Bellucci, & Bagnoli, 2017). Como resultado, la información que se posee acerca de las redes de transporte público ha mejorado sustancialmente (Moya-Gómez et al., 2017).

Twitter es una fuente de datos muy valiosa para extraer opiniones y sentimientos sobre multitud de temas. En la actualidad, millones de personas comparten opiniones y sentimientos sobre diferentes aspectos de la vida real en las redes sociales, lo que las convierten en recursos ricos de datos para minería de opinión y análisis de sentimientos (Kocich, 2017). Hay un rango amplio de investigaciones en varios campos que usan el texto incluido en los *tweets* para analizar sentimientos y obtener resultados valiosos para las empresas. El contenido semántico de un *tweet* es difícil de interpretar y analizar, pero recientemente han habido avances en técnicas de minería de texto aplicadas a *tweets* para interpretar cuantitativamente su contenido (Lansley & Longley, 2016). Además de ser fuentes de datos de muy bajo coste y que pueden ser recopilados a tiempo real, el uso de datos de redes sociales como *Twitter* sobre las encuestas tradicionales de movilidad tiene ventajas como ser capaz de observar las necesidades específicas sobre un tema, o de obtener información acerca de los motivos detrás de un sentimiento particular (Collins et al., 2013). Sin embargo, el contenido de los datos semánticos de *Twitter* permanece descontextualizado en este tipo de estudios, a no ser que se encuentren métodos para atarlo al campo de la geografía (Graham et al., 2014). A pesar de que el uso de datos de *Twitter* para la minería de opinión es popular en varios campos, su uso en la administración, desarrollo y planteamiento del transporte aún están limitados (Luong & Houston, 2015).

Este caso de estudio tiene como objetivo explorar las percepciones de los usuarios de *Twitter* cuando viajan en un sistema de transporte público como el Metro de Madrid. Sin

embargo, los *tweets* geolocalizados corresponden al 1% de los datos de *Twitter* que se pueden descargar a partir de la API que ofrece la red social. Además, a la hora de estudiar un tema en concreto, las muestras son muy pequeñas y presentan un ruido considerable (Graham et al., 2014). En este caso de estudio, se propone usar el texto de *tweets* no geolocalizados para extraer información espacial a partir de la referencia artificial o geocodificación de palabras claves. Como se explicará en el siguiente apartado, esta aproximación permite obtener una muestra mayor y con menor ruido. Aun así, es importante poder geolocalizar estos datos de *Twitter* ya que la localización de las paradas o estaciones de un servicio de transporte, y la integración de estas paradas a la red de transporte público urbano son factores claves para poder estudiar la calidad global de la conexión del servicio (Moyano, Moya-Gómez, & Gutiérrez, 2018). A continuación, se pueden extraer los temas más relevantes y los sentimientos de los usuarios del sistema de transporte público de Metro de Madrid, y ubicarlos en el espacio. Además, se propone un modelo de Regresión Geográficamente Ponderada (GWR) con el objetivo de explorar la causalidad de las variables espaciales que afectan al número de usuarios con sentimientos negativos. Para ello se usarán datos de fuentes oficiales que permitan usar como variables exploratorias la población, el número de puntos de interés o de líneas de autobús próximas a las estaciones de metro.

4.5.2. Metodología específica para la extracción de temas y sentimientos de los textos de Twitter

La razón principal por la que en este capítulo se trabaja con *tweets* no geolocalizados (menciones a la cuenta pública de *Twitter* del Metro de Madrid) radica en la hipótesis de que los usuarios de transporte público tienden a contestar directamente a la cuenta oficial de las agencias de transporte público cuando se quejan o alaban un servicio ofrecido, o cuando detectan un fallo en una localización determinada. De este modo, la frecuencia de datos es mucho mayor que una muestra de *tweets* normales, y los textos tienden a dar una visión sobre quejas muy específicas o información relacionada con el servicio, mientras que una muestra normal de *tweets* tiende a escribir sobre temas comunes (Haghighi et al., 2018). Se ha seleccionado el Metro de Madrid como caso de estudio al ser el servicio de transporte público que recibe mayor número de mensajes de los usuarios de *Twitter* (Tabla 22).

Tabla 22: Frecuencia de *tweets* de cuentas relacionadas con sistemas de transporte público del Área Metropolitana de Madrid durante una semana (16 a 22 de septiembre).

Cuenta de usuario	**Servicio**	**Escala**	**Número de *tweets* respondiendo a la cuenta**
@metro_madrid	Metro	Urbana-Metropolitana	4.004
@CercaniasMadrid	Train	Metropolitana	2.686
@EMTMadrid	Bus	Urbana	769

Fuente: Elaboración propia a partir de datos de *Twitter*.

El análisis semántico de los *tweets* conlleva un desafío ya que es necesario trabajar con mensajes desestructurados, muy cortos, que contienen ruido y errores como lenguaje no común, abreviaciones, símbolos, faltas de ortografía, o acrónimos (Haghighi et al., 2018; Hiltz et al., 2014). Para poder obtener información relevante de los textos de los *tweets*, hace falta realizar una serie de pasos de preprocesado, limpieza, y preparación. Para ello, se ha empleado la librería *Pandas* incluida en el lenguaje *Python* (Tabla 23).

Tabla 23: Pasos para la limpieza de texto de los *tweets.*

Orden	**Actividad**
1	Conversión de letras a minúsculas
2	Transformación de caracteres especiales del idioma español (transformación de caracteres á/é/í/ó/ú/ü/ñ en caracteres a/e/i/o/u/u/n).
3	Eliminación de hipervínculos (“palabras” que empiezan con “http” o “https”).
4	Eliminación de menciones (palabras que empiezan con el carácter @).
5	Eliminación de caracteres espaciales (por ejemplo, caracteres #, /, o –).
6	Eliminación de signos de puntuación.
7	Eliminación de *stopwords* (palabras sin significado como artículos o preposiciones).
8	Tokenización (división del texto de cada *tweet* por las palabras contenidas en el texto).

Fuente: Elaboración propia.

Al usar *tweets* no geolocalizados, se obtiene una muestra mayor de datos en menor tiempo que descargando *tweets* geolocalizados, pero la muestra carece de coordenadas de

longitud y latitud, por lo que los *tweets* no pueden ser localizados en el espacio. La solución propuesta para este problema es extraer el nombre de paradas de Metro de los textos de los *tweets*, bajo la hipótesis de que los usuarios que viajan en transporte público tienden a usar un vocabulario muy específico de localización, escribiendo frecuentemente la estación o línea donde se encuentran (Haghighi et al., 2018).

Con este objetivo, se elaboró un diccionario de palabras claves escritas en *Python* con el nombre de todas las estaciones de Metro y las líneas en las que están incluidas. Adicionalmente, se realizó una búsqueda de abreviaciones potenciales para que fuesen reemplazadas por el nombre completo de la estación. Por ejemplo, se elaboró una búsqueda en el campo de texto de los *tweets* con palabras que incluían la palabra *ppio*, y los *tweets* encontrados fueron georreferenciados con las coordenadas espaciales de la estación de Metro de Príncipe Pio. De este modo, se geocodificaron 3.454 *tweets* pertenecientes a 2.418 usuarios (el 12,5% de la muestra inicial). Para los mensajes geocodificados en una estación con varias líneas de metro, se utilizó un segundo diccionario de palabras con el nombre de las líneas de Metro, con el objetivo de intentar extraer una única línea de Metro del texto de los *tweets*. Aunque la geocodificación es una solución para añadir información espacial a los *tweets* no geolocalizados, el porcentaje de *tweets* que se geocodifican es bajo. Hay que considerar que los *tweets* no geocodificados normalmente contienen quejas más comunes sobre el servicio, que no relacionadas con ninguna estación concreta.

Para poder analizar el contenido semántico generado por los usuarios de *Twitter*, una vez geocodificados los *tweets* se eliminaron las palabras cortas de los textos (palabras con 3 o menos caracteres) y los caracteres numéricos. Tomando como ventaja la homogeneidad de la temática de los textos de los *tweets*, se formularon una serie de temas basados en las principales quejas que el sistema de Metro recibe en las redes sociales. A continuación, se elaboró un diccionario de *Python* que contiene una colección de palabras claves específicas para cada tema, con el objetivo de identificar el tema principal de cada *tweet* (Tabla 24). Con este método se encontró que sobre el 50% de los *tweets* de la muestra contiene un tema. Sin embargo, debido a las limitaciones del diccionario, no se halló el tema de una importante cantidad de *tweets* que si están potencialmente relacionados a un tema.

Tabla 24: Lista y descripción de los temas formulados.

Tema	**Definición**	**Palabras clave**
Puntualidad	*Tweets* que abordan quejas sobre la frecuencia o lentitud del servicio.	tarde, lento, retraso, frecuencia, esperando, etc.
Confort	*Tweets* que tratan temas sobre el bienestar de los usuarios como temperatura, limpieza, o seguridad.	ventilacion, calor, sucio, huele, asfixiados, etc.
Averías	*Tweets* que reportan fallos o averías del sistema.	averia, obras, interrumpido, fallo, suspendido, etc.
Sobresaturación	*Tweets* que expresan problemas de sobresaturación en las estaciones o trenes.	lleno, saturados, aglomeracion, caber, hacinados, etc.

Fuente: Elaboración propia a partir de datos de *Twitter*.

Estos *tweets* fueron agrupados por grupos mediante un modelo *Latent Dirichlet Allocation* (LDA) (Blei, Ng, & Edu, 2003). Este modelo matemático es usado para hallar temas a partir de la frecuencia de palabras de una serie de documentos, donde un tema es representado como una lista de palabras. Para ello, divide una serie de muestras en temas, y mide la frecuencia de las palabras y el número de veces que se repiten en la base de datos. Estas palabras son divididas en clústeres, y los clústeres son nombrados tras visualizar las palabras más usadas combinadas en una frase con sentido (Saura & Bennett, 2019). El modelo LDA usado en esta investigación se realizó con la librería LDA 1.0.5. usando la muestra de *Gibbs*. Este algoritmo está disponible en una biblioteca gratuita de Python, llamada *Gensim*. En este trabajo, se han establecido cinco temas para la implementación del modelo LDA. Cuatro clústeres fueron relacionados con los cuatro temas previamente establecidos con una precisión de casi el 70%. El quinto clúster, compuesto por 298 *tweets*, trata temas varios diferentes a los cuatro temas definidos.

A continuación, se elaboró un análisis de sentimientos para asignar un valor positivo o negativo a cada uno de los *tweets* de la muestra. Para este propósito, se utilizó un modelo de aprendizaje profundo *BERT* (*Bidirectional Encoder Representations from Transformers*). Este modelo fue entrenado para el análisis de sentimientos de *tweets* en idioma español a partir de dos muestras de *tweets*. Para comprobar su efectividad, se realizó un experimento que mezcló ambas muestras, y entonces las dividió en una muestra de entrenamiento (80% de los *tweets*), y una muestra de prueba (el 20% restante). Los resultados de los experimentos pueden ser vistos en la Tabla 25. Los valores de *F1-Score* y de exactitud están sobre el 0,90, indicando que el modelo de entrenamiento puede

evaluar con bastante precisión el sentimiento de los *tweets*. Este modelo fue utilizado para asignar un valor de 0 a los *tweets* negativos, y valor de 1 a los *tweets* positivos. Finalmente, se han agrupado los datos espacialmente a nivel de líneas y estaciones de metro, y temporalmente por día y número de hora.

Tabla 25: Parámetros de evaluación del modelo de entrenamiento de sentimiento.

Parámetros de evaluación	**Muestra de entrenamiento**
Precisión	0,95
Recuerdo	0,87
Exactitud	0,90
F1-Score	0,91

Fuente: Elaboración propia.

En cuanto al estudio de las variables que permitan discernir las causas tras la distribución espacial de los comentarios realizados por los usuarios de *Twitter* y sus temas, se ha realizado primero un Análisis Exploratorio de los Datos, que incluye análisis de distribución de cada variable. A continuación, se preparó un modelo OLS para evaluar las relaciones globales y el comportamiento de las variables. A partir de los resultados obtenidos en el modelo de regresión lineal OLS, se realizó un análisis de autocorrelación espacial Moran I para comprobar si la distribución espacial de los residuos era aleatoria o estaba clusterizada. Al comprobar que los valores de los residuos estaban geográficamente clusterizados y que el modelo OLS no contaba con un suficiente poder explicativo, se procedió a realizar un modelo GWR.

El modelo GWR (Brunsdon, Fotheringham, & Charlton, 1996) fue desarrollado en respuesta a las condiciones heterogéneas con relaciones espaciales variables en el caso de que los modelos de regresión espaciales globales fallen. En este modelo, los coeficientes β_n en los predictores x_n pueden variar espacialmente en unas coordenadas geográficas $(u_i v_i)$ de una localización i. Por cada localización i, el valor de la variable dependiente y_i es estimada como indica la siguiente ecuación:

$$y_i = \beta_0(u_i v_i) + \beta_1(u_i v_i)\, x_1 + \beta_2(u_i v_i)\, x_2 + \ldots + \beta_n(u_i v_i)\, x_n + \varepsilon_i$$

Las ventajas principales del modelo GWR incluyen una mayor precisión, la reducción del problema de la autocorrelación espacial de los modelos OLS y la variación de los coeficientes a través del espacio. El modelo GWR permite investigar patrones espaciales

de estimaciones locales y analizar posibles causas, reconocer donde tienen las variables independientes un mayor o menor poder de explicación, y apoyar las políticas de toma de decisiones a nivel local (Cardozo, García-Palomares, & Gutiérrez, 2012).

La Tabla 26 indica las variables usadas tanto en los modelos OLS como GWR. Estas variables pueden ayudar a explicar el número de quejas, al hipotetizar un mayor número de comentarios de usuarios en las estaciones cercanas a los puestos de trabajo, puntos de interés, y conexiones con otros servicios de transporte. Un modelo OLS puede ser de poca utilidad en esta situación ya que no interpreta como afectan las variables a la distribución espacial de los comentarios. En cambio, un modelo GWR considera la variación espacial que pueden ejercer las variables sobre la muestra (por ejemplo, el número de residentes cercanos a una estación o las conexiones de la red de metro con otros servicios pueden generar un mayor número de quejas en un sitio o en otro) y obtiene coeficientes locales que reflejan más exactamente su influencia en el número de comentarios sobre el metro (Cardozo et al., 2012). Por ejemplo, es posible detectar coeficientes más altos en la variable de número de residentes en zonas con un menor nivel de renta (lugares más dependientes del transporte público, por lo que sus habitantes pueden ser más propensos a sufrir los problemas hallados en la red).

Tabla 26: Variables utilizadas en los modelos OLS y GWR.

Variable	**Tipo**	**Fuente**
Número de usuarios de *Twitter* con sentimiento negativo	Dependiente	*Twitter*
Número de residentes en edad de trabajar (radio 500 metros)	Exploratoria	Censo 2019 Instituto Nacional Estadística
Número de puntos de interés (radio 500 metros)	Exploratoria	Datos de 2019 de *OpenStreetMaps*
Número de conexiones con otros servicios de transporte (otras líneas de metro, Cercanías, etc.)	Exploratoria	Consorcio de Transportes de Madrid
Número de líneas de autobuses EMT, urbano e interurbano (radio 500 metros)	Exploratoria	Ficheros *GTFS* Consorcio de Transportes de Madrid

Fuente: Elaboración propia a partir de datos de *Twitter*.

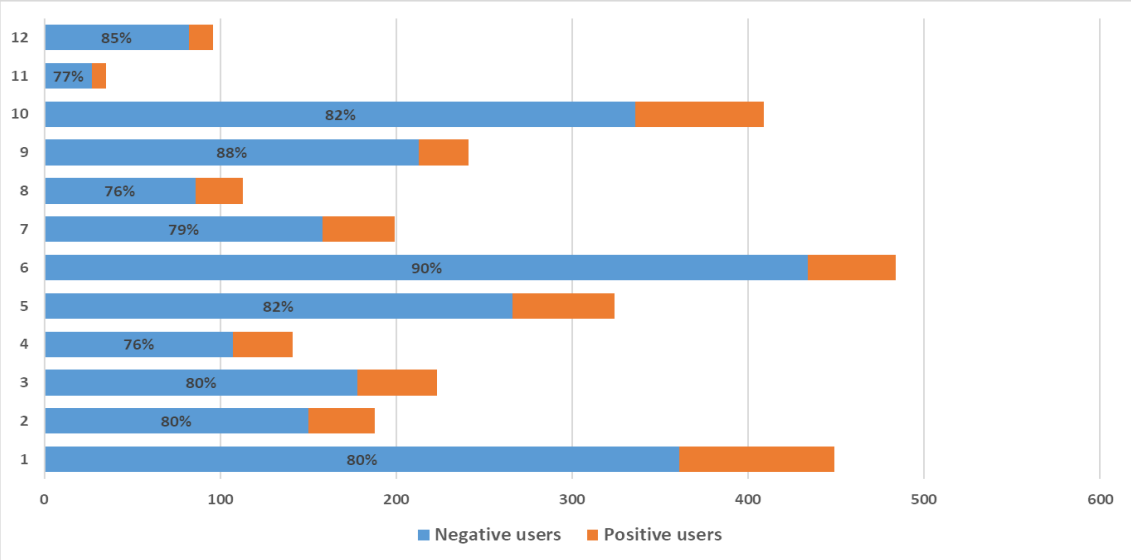
El modelo GWR fue efectuado a resolución de estaciones de metro, con un total de 241 observaciones). Se seleccionó un modelo de 14 vecinos próximos al ser el número que proporcionaba el valor AICc más reducido. Este modelo está basado en la optimización de los indicadores estadísticos seleccionados (R^2, AICc, residuos cuadrados, *sigma*), el testeo de la normalidad y la autocorrelación espacial de errores, la comprobación de residuos estandarizados dentro de un umbral de 2,5 desviaciones estándar, la distribución de los valores r^2, y la contribución explicativa de las variables independientes.

4.5.3. Distribución espacio-temporal de los usuarios de Twitter con sentimiento negativo

De los 2.418 usuarios de *Twitter* encontrados, 2.097 usuarios han publicado algún *tweet* con sentimiento negativo, mientras que solo 475 usuarios han publicado algún *tweet* positivo. Se puede decir que estos primeros resultados cumplen con la hipótesis que sugiere que los usuarios de *Twitter* elaboran principalmente quejas cuando interactúan con la cuenta oficial del servicio de transporte público. Las líneas de Metro 6, 9 y 12 presentan un mayor porcentaje de usuarios con sentimientos negativos, mientras que las líneas 4, 8 y 11 cuentan con el menor porcentaje negativo de usuarios. Las líneas más frecuentadas por los usuarios de *Twitter* detectados son la línea 6 (línea circular que rodea el área central de Madrid y que incluye todas las estaciones que sirven como intercambiadores de autobús), la línea 1 (línea que conecta el centro de la ciudad con las estaciones de tren y los lugares de oficinas y negocios del norte de la ciudad, y con los espacios residenciales del sur de la ciudad), y la línea 10 (línea que cruza toda el área metropolitana de norte a sur) (Figura 53).

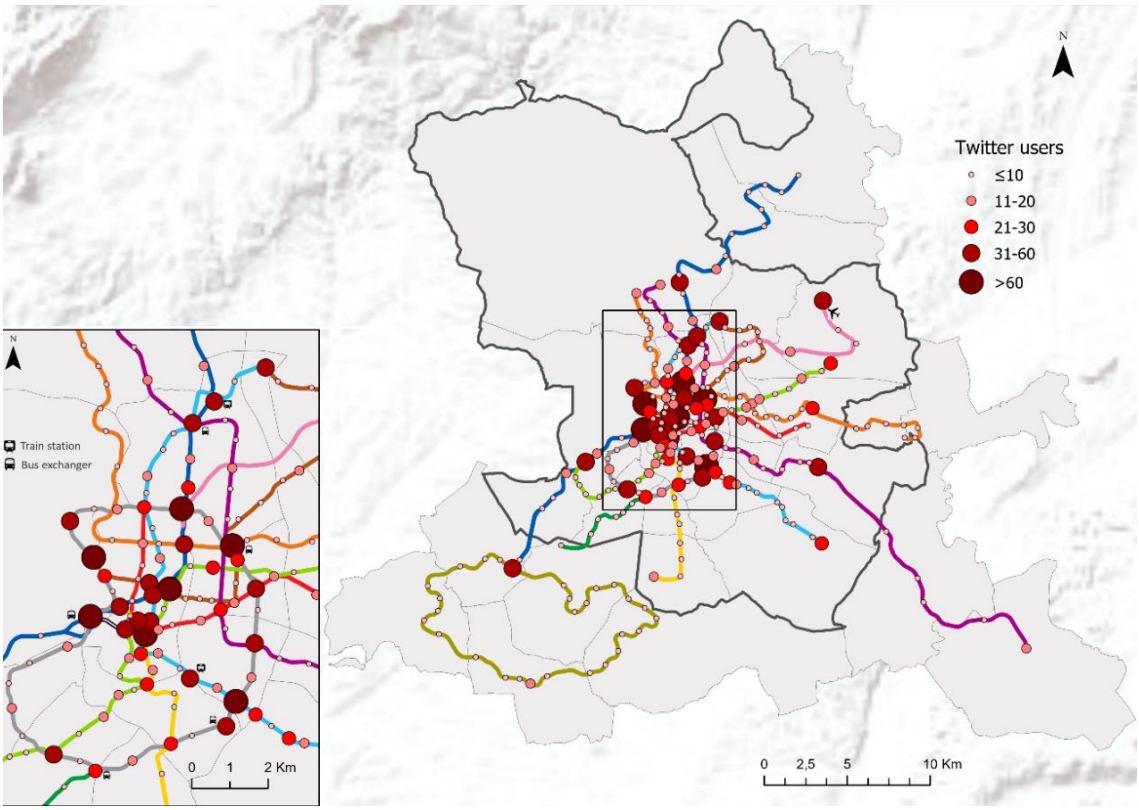
Las estaciones de Metro más frecuentes en la muestra corresponden con las más transitadas de acuerdo con los datos oficiales: las estaciones localizadas en las zonas centrales de la ciudad de Madrid, y las estaciones de la línea 6 que conectan diferentes líneas de Metro entre sí y con otros servicios de transporte público como autobuses. También se pueden destacar estaciones localizadas en zonas periféricas que cuentan con servicios importantes (como las correspondientes a las dos estaciones de trenes de Atocha y de Chamartín, o la estación que transborda con el aeropuerto de Barajas). Otras estaciones con un peso destacado son los intercambiadores con el servicio de Metro en los municipios periféricos (Figura 54).

Figura 53: Número de usuarios de *Twitter* totales y porcentaje por sentimientos por línea de Metro.



Fuente: Elaboración propia a partir de datos de *Twitter*.

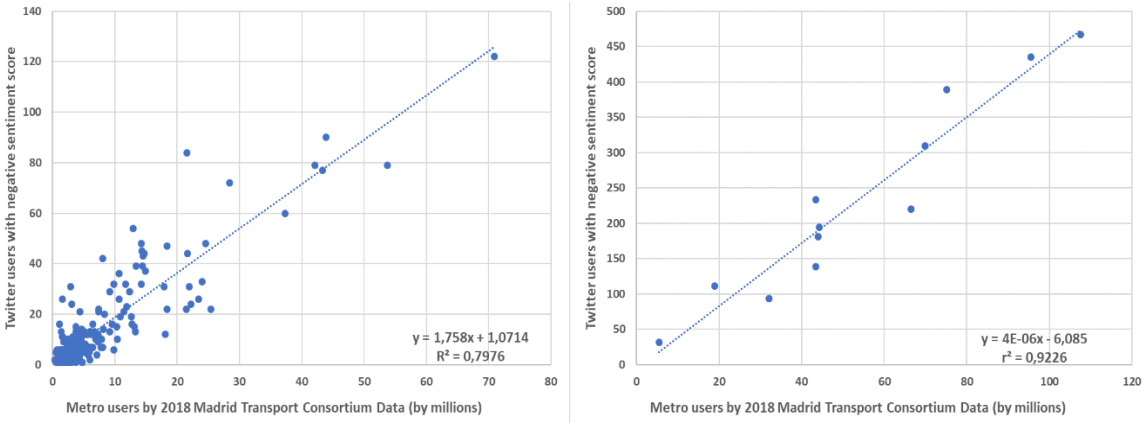
Figura 54: Distribución de usuarios de *Twitter* con sentimiento negativo en la red de Metro de Madrid.



Fuente: Elaboración propia a partir de datos de *Twitter*.

El coeficiente r^2 de los usuarios de *Twitter* por estación de Metro con los datos oficiales del Consorcio de Transportes de Madrid marca un valor de 0,80, mostrando una distribución de usuarios cercana a la realidad que recogen los datos oficiales. El valor del coeficiente aumenta a 0,92 si se agregan los usuarios por línea de Metro y se compara el valor con los datos oficiales de número de viajeros por línea, lo que indica una mayor precisión. En esta comparación, hay estaciones con una sobreestimación de usuarios de *Twitter* respecto a la realidad, principalmente en estaciones de la línea 10 o en estaciones ubicadas en la zona central. Dos estaciones (Pacífico en las líneas 1 y 6, y Bilbao en las líneas 1 y 4) destacan por su sobrestimación de usuarios en la muestra. Mientras, se da una subestimación de usuarios de *Twitter* respecto a los datos oficiales en estaciones pertenecientes a la línea 4 de metro, en las estaciones localizadas en los distritos del sur del municipio (línea 3), y en las estaciones que se ubican en los municipios adyacentes (principalmente en las estaciones del cinturón sur conectadas por la línea 12) (Figura 55).

Figura 55: Relación entre número de usuarios de *Twitter* con sentimiento negativo y número de viajeros registrados en datos oficiales a nivel de estaciones (izquierda) y líneas de Metro (derecha).

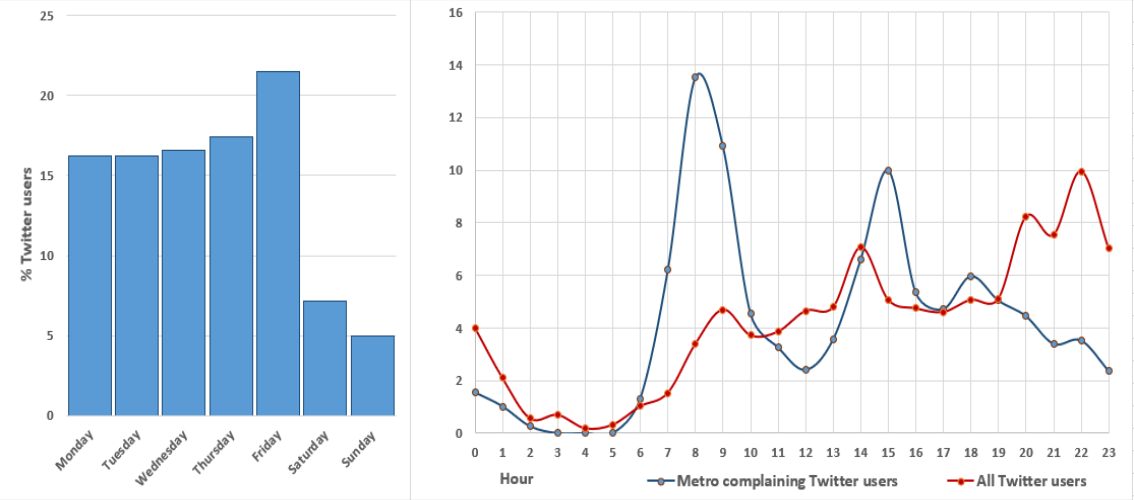


Fuente: Elaboración propia a partir de datos de *Twitter* y de datos del Consorcio de Transportes de Madrid de 2018.

Durante la semana, hay un porcentaje constante y uniforme de usuarios durante los días de trabajo, siendo el viernes el día con mayor número de usuarios. Esta situación contrasta con los fines de semana, donde el porcentaje de usuarios de *Twitter* baja drásticamente. Se puede asumir que la principal razón reposa en el motivo de viaje, ya que el principal uso del Metro es viajar al trabajo o lugar de estudios. Esta casuística también se refleja analizando el número de usuarios por hora. Se observan dos momentos cumbre: un pico mayor por la mañana (8 horas) cuando la población viaja al trabajo o a estudiar, y un pico

menor a primera hora de la tarde (15 horas) cuando la gente regresa a casa. En contraste, hay un bajo uso del Metro durante el mediodía, ya que los ciudadanos se hallan en sus puestos de trabajo o estudio, por lo que la movilidad metropolitana desciende. Se puede observar también un constante descenso de la actividad por la tarde (desde las 18 horas) que se prorroga durante la noche, paralelamente a la disminución del servicio por las noches. Este perfil temporal contrasta con la distribución temporal de una muestra de un día de usuarios de *Twitter* no relacionados con temáticas de transporte. Estos usuarios presentan poca actividad por la mañana y tienden a publicar mensajes por la noche, en contraposición con el perfil de usuario de *Twitter* que reporta problemas al servicio principalmente por la mañana (Figura 56).

Figura 56: Porcentaje de usuarios de *Twitter* con sentimiento negativo en el Metro de Madrid por día y hora.



Fuente: Elaboración propia a partir de datos de *Twitter*.

4.5.4. Distribución espacio-temporal de los principales temas reportados

El principal problema reportado por los usuarios de *Twitter* que utilizaron el servicio de Metro es la puntualidad: Mientras, la sobresaturación es el tema con menor número de usuarios, pero también presenta el mayor porcentaje de usuarios sobre el total (Tabla 27).

La Figura 57 muestra la distribución de los principales temas con sentimiento negativo en cada estación de metro. Se pueden visualizar algunos patrones. Por ejemplo, las averías son el tema principal con sentimiento negativo en cada estación ubicada en el segundo segmento de la línea 1, o en la parte este de la línea 12, mientras que las estaciones de la

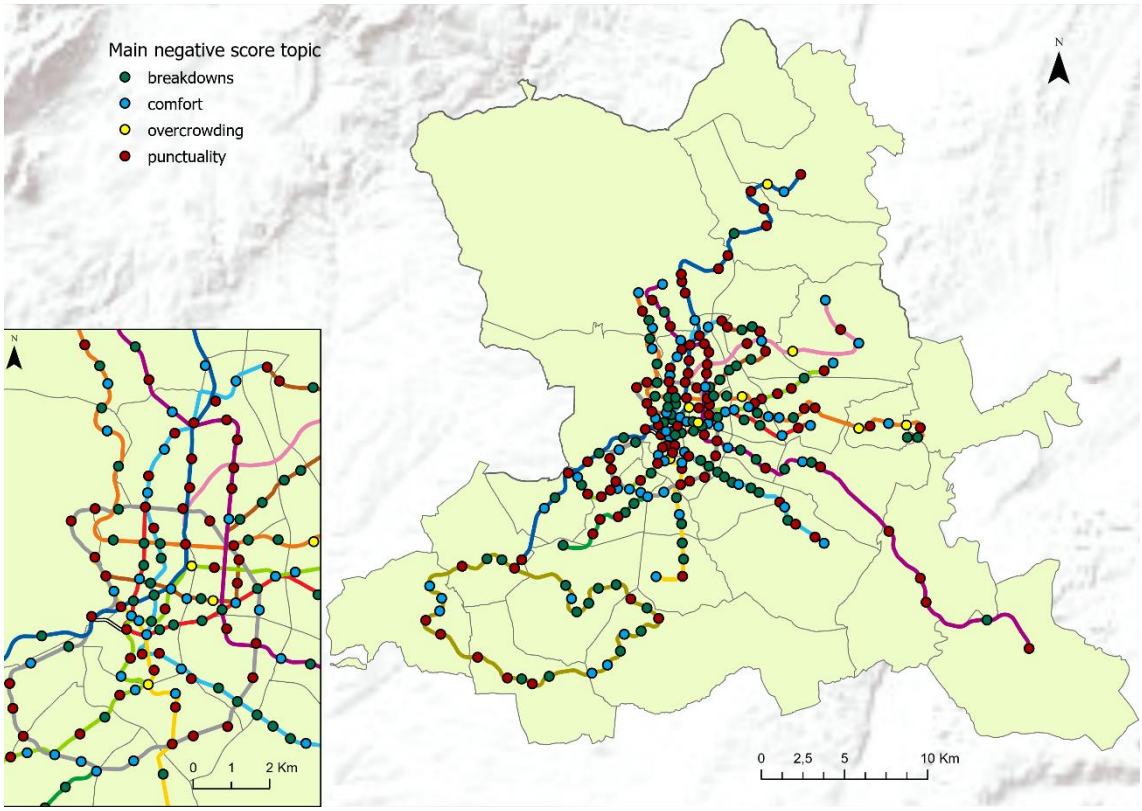
línea 9 están afectadas por problemas de puntualidad. La puntualidad y las averías son los temas principales más distribuidos a lo largo de la red, con mayor visibilidad de la puntualidad en las estaciones de la parte central de la ciudad, mientras que las averías son más visibles en las estaciones de la periferia. También se aprecia una mayor visibilidad de los problemas de confort en los distritos centrales de la ciudad de Madrid.

Tabla 27: Número de usuarios de *Twitter* por tema.

Tema	**Número de usuarios**	**Número de usuarios con sentimiento negativo**	**% usuarios con sentimiento negativo**
Puntualidad	1016	859	84,54%
Confort	729	555	76,13%
Averías	809	666	82,32%
Sobresaturación	372	345	92,74%

Fuente: Elaboración propia a partir de datos de *Twitter*.

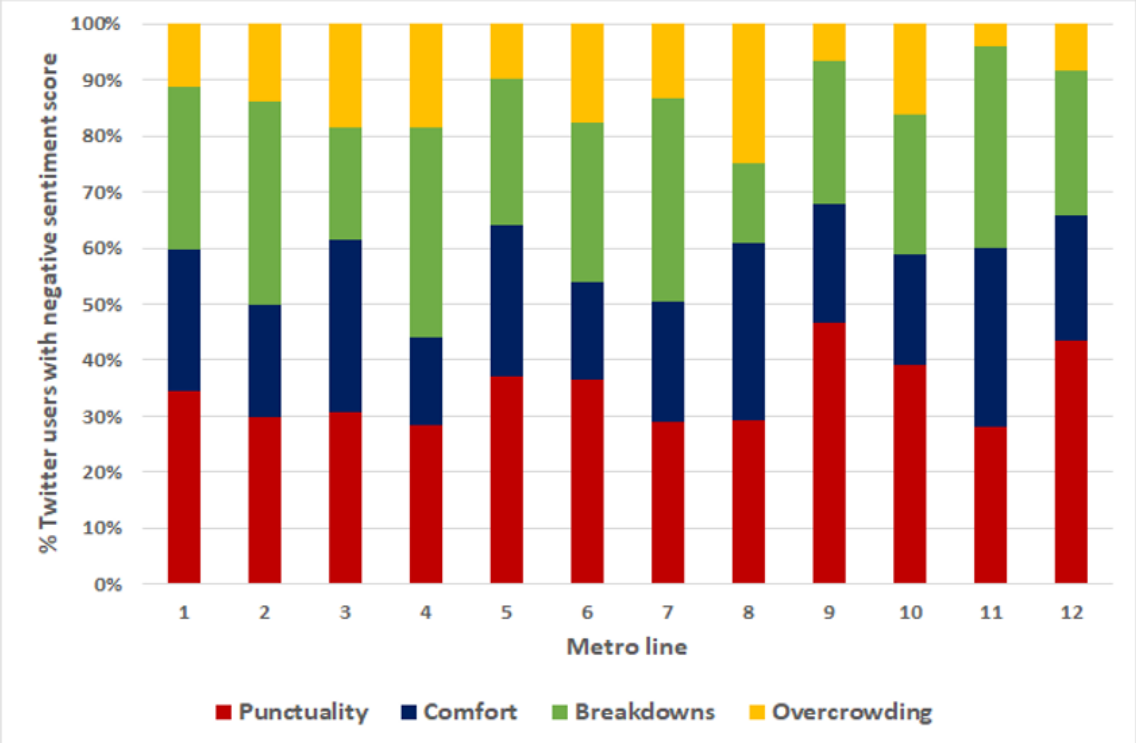
Figura 57: Principal tema con sentimiento negativo en las estaciones del Metro de Madrid.



Fuente: Elaboración propia a partir de datos de *Twitter*.

Las líneas de Metro más utilizadas (líneas 1, 6 y 10) son también las que presentan un mayor número de usuarios que reportan problemas. Sin embargo, las quejas pueden ser diferentes en cada línea. La puntualidad es un problema importante en casi todas las líneas de metro, además de ser la queja más reportada en las líneas de Metro más utilizadas. Pero las líneas que muestran un mayor porcentaje de usuarios con sentimiento negativo en este tema corresponden con las líneas 9 y 12. El confort es un problema de interés en las líneas largas con muchas estaciones (líneas 1, 5, y 6), pero es principalmente visible en la línea 8 (la línea más corta de la red, diseñada especialmente para conectar la ciudad con el aeropuerto). Las líneas que alcanzan la periferia de la ciudad (líneas 2, 4, 7 y 11) muestran principalmente problemas de averías. La sobresaturación destaca en las líneas con muchas estaciones que conectan con otras líneas de Metro (líneas 6 y 10) pero también en la línea 8 (Figura 58).

Figura 58: Porcentaje de usuarios de *Twitter* con sentimiento negativo por tema y línea.

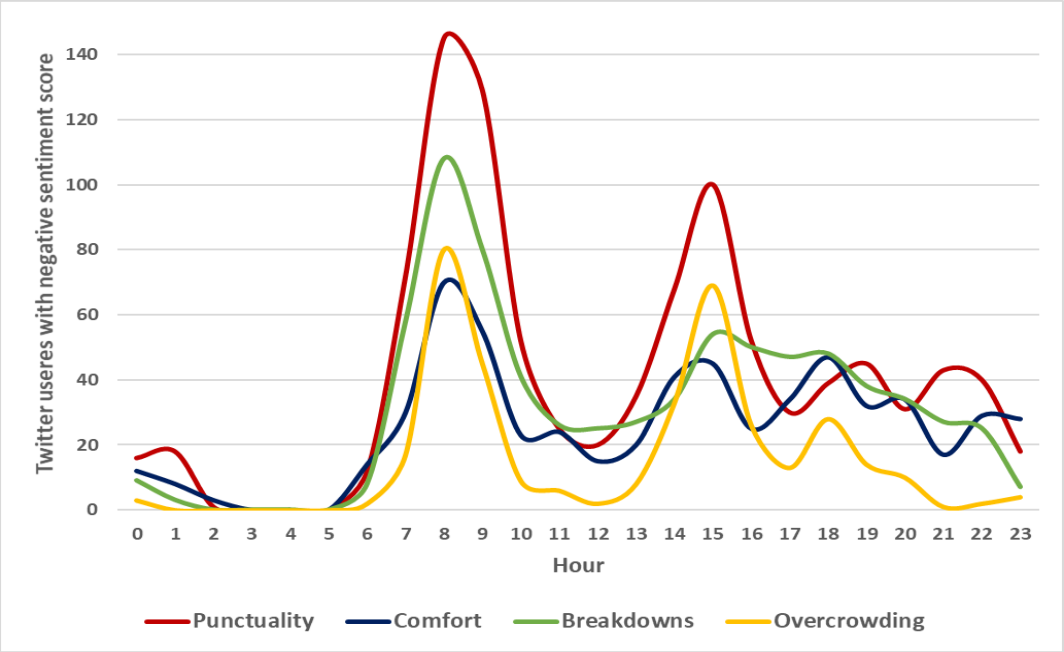


Fuente: Elaboración propia a partir de datos de *Twitter*.

Durante el día, los cuatro temas presentan un máximo de usuarios con sentimiento negativo durante los dos momentos cumbre del día (mañana y primera hora de la tarde). La puntualidad es de nuevo el tema principal durante casi todo el día, siendo claramente visible en estos dos picos. Se puede apreciar una mayor cantidad de usuarios en el pico de la mañana, donde la puntualidad tiene una diferencia relativa mayor respecto a otros

temas. El perfil temporal de la sobresaturación es similar, con la diferencia de que muestra un número casi igual de usuarios en los dos momentos pico (como consecuencia, la diferencia relativa respecto a los problemas de puntualidad es menor en el segundo pico). A pesar de ser el tema principal durante una parte de la tarde, las averías presentan un mayor número de usuarios reportando quejas por la mañana. Esta situación es similar con el confort, a pesar de ser también prominente a mediodía y por la tarde (Figura 59).

Figura 59: Número de usuarios de *Twitter* con sentimiento negativo en cada tema por día y hora.



Fuente: Elaboración propia a partir de datos de *Twitter*.

4.5.5. Análisis de la causalidad espacial (GWR)

Al analizar los resultados del modelo GWR aplicado a la densidad de usuarios con sentimiento negativo, el valor R^2 presenta valores más elevados que el modelo OLS. El valor AIC tiene valores menores que los presentados por el modelo OLS, mostrando una mayor capacidad explicativa. Además, el análisis Moran I revela que el modelo OLS presenta clústeres en los residuos, que se corrigen en parte en el modelo GWR (Tabla 28).

Tabla 28: Estadísticas de los diagnósticos OLS y GWR.

Modelo	**R^2**	**R^2 ajustado**	**AICc**	**Índice Moran**
OLS	0,54	0,53	1873,40	0,03
GWR	0,71	0,64	1836,86	-0,07

Fuente: Elaboración propia.

Los residuos muestran un buen ajuste entre el número de usuarios con sentimientos negativos y las variables analizadas, aunque se puede encontrar una subestimación de las quejas de usuarios en las estaciones del centro de Madrid. Los valores de R^2 locales indican una buena precisión del modelo en el norte de la zona de estudio, mientras que los valores más bajos se sitúan en las estaciones de la línea 1 ubicadas en los barrios de Vallecas, las estaciones de la periferia este de la línea 7 y las estaciones de la línea 12 (Figura 60).

La variable de población en edad de trabajar tiene coeficientes con valores negativos en casi toda la red de transporte, lo que puede señalar que los usuarios de Metro principalmente reportan problemas al llegar al lugar de trabajo y raramente expresan quejas en las estaciones en las que acceden a la red. Sin embargo, esta variable presenta signos positivos en las estaciones ubicadas en el este y sureste del área metropolitana y en el municipio de Getafe. Se trata de áreas residenciales habitadas por trabajadores con un nivel medio de renta, dependientes del transporte público, que lo usan mucho y en desplazamientos más largos. Los coeficientes son más bajos en el centro de la red, donde la utilidad de esta aumenta, los viajes de este tipo de residentes tienden a ser más cortos, y es más probable que los usuarios que reporten mensajes puedan ser turistas.

El número de puntos de interés presenta coeficientes positivos en todas las estaciones, de manera que un mayor número de puntos de interés produce un aumento del número de usuarios de *Twitter* y por tanto más quejas. Este resultado concuerda con los coeficientes obtenidos en la variable anterior, ya que hay un mayor número de puntos de interés en las zonas de trabajo o servicios específicos que en los lugares residenciales. Sin embargo, este incremento es variable en el espacio, con elasticidades más altas en el centro, norte y sudoeste de la ciudad, zonas con un gran número de infraestructuras y servicios, que es posible que atraigan viajes más largos. En cambio, esta variable presenta una baja

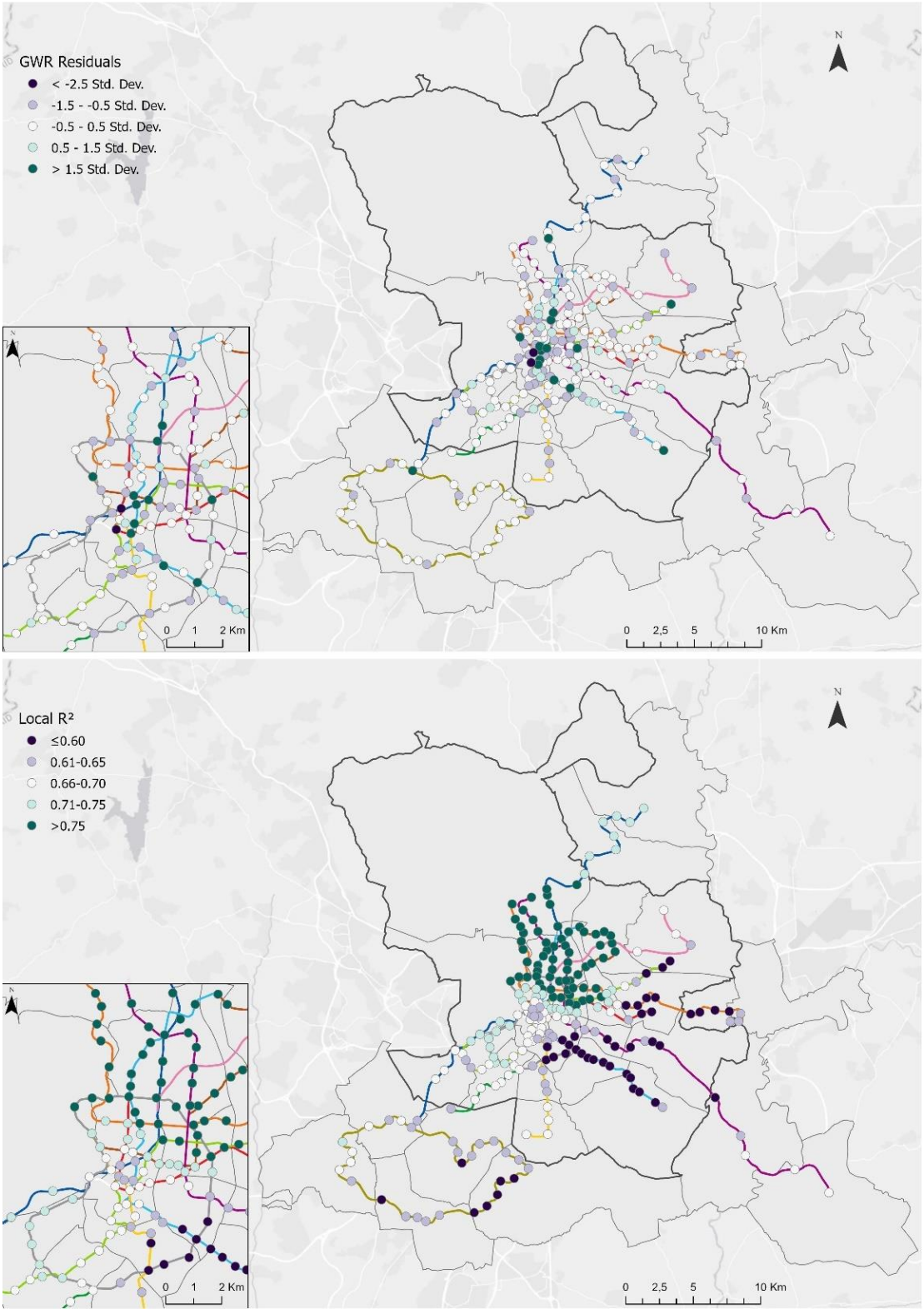
elasticidad en las zonas periféricas del este y sudeste de Madrid y en algunos municipios del sur del área metropolitana.

El coeficiente de número de conexiones con otros servicios de transporte público también presenta coeficientes positivos en todas las estaciones de metro. La elasticidad es alta en las estaciones del centro y norte de Madrid, ya que son las áreas con mayor número de estaciones intermodales (que como se ha visto, son las estaciones con mayor número de usuarios reportando problemas), donde las conexiones generan mayor número de intercambios y donde las estaciones son más antiguas e incómodas. Mientras, las estaciones localizadas en las áreas sur y suroeste de la ciudad, el distrito de Barajas y los municipios periféricos presentan menor elasticidad, lo que indica un menor número de usuarios en las zonas donde la conectividad de la red es menor y muchas veces las estaciones más modernas y cómodas para el intercambio entre líneas.

Por último, el número de líneas de autobuses próximas presenta valores negativos en el centro y sur del área metropolitana, pero cuenta con coeficientes positivos en las estaciones localizadas en el norte y oeste del área de estudio. Este resultado puede indicar un mayor número de desplazamientos combinados en estas áreas, que destacan por poseer un nivel de renta elevado, ser zonas orientadas a actividades de trabajo o servicios específicos y con un menor número de estaciones de metro que conecten de forma directa a esos puestos de trabajo o servicios (Figura 61).

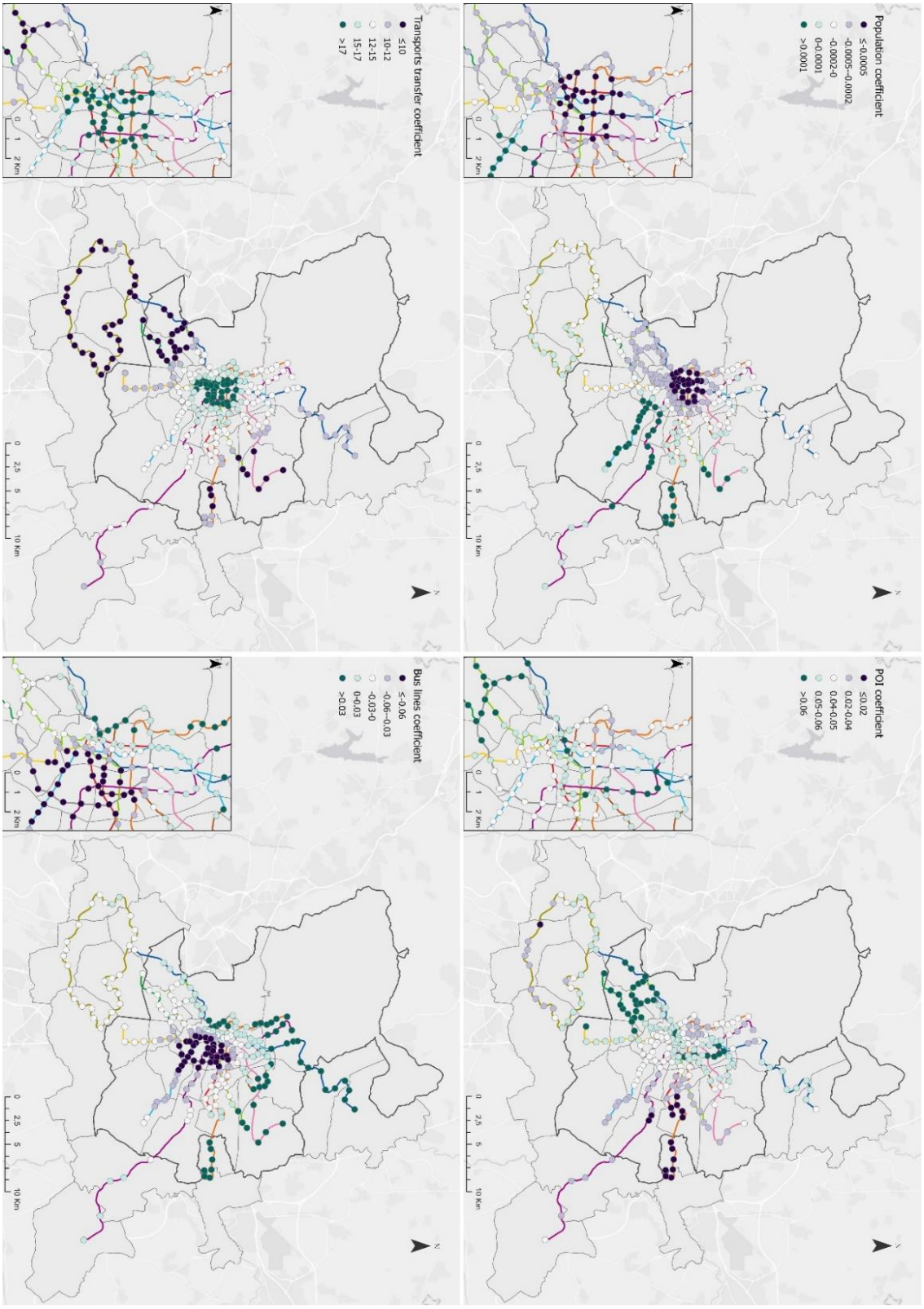
Aunque los resultados obtenidos al analizar los coeficientes espaciales de cada variable pueden servir de apoyo para observar que zonas son más propensas a ser afectadas por cada variable, hay que tener en cuenta que, aunque el modelo presenta coeficientes aceptables, estos coeficientes no son muy elevados (tal como se recoge en la Tabla 2, el valor R^2 del modelo es de 0,71, y el R^2 ajustado es de 0,64). Finalmente, el modelo GWR es un modelo exploratorio, por lo que las interpretaciones de los coeficientes son difíciles de explicar y su utilidad es relativa.

Figura 60: Distribución residuos y valores R^2 locales en el Área Metropolitana de Madrid.



Fuente: Elaboración propia a partir de datos de *Twitter*.

Figura 61: Distribución de los coeficientes de las variables exploratorias utilizadas en el modelo GWR en el Área Metropolitana de Madrid.



Fuente: Elaboración propia a partir de datos de *Twitter*.

5. CONCLUSIONES

El último capítulo de la tesis doctoral establece las conclusiones de la investigación en forma de respuesta a las preguntas de investigación formuladas en el apartado 1.2. de la tesis. Además, se elaboran unas conclusiones generales con las que se responde al objetivo principal de la tesis. Finalmente, este capítulo introduce toda una serie de futuras líneas de investigación que se han abierto durante el desarrollo de la investigación, y que podrán ser desarrolladas en la etapa postdoctoral.

5.1. Respuestas a las preguntas de investigación

a) ¿Las fuentes de datos generadas por las TIC son adecuadas para el estudio de la movilidad urbana?

Las fuentes de datos tradicionales cuentan con información rica y valiosa sobre la movilidad urbana, pero su preparación conlleva de un coste y tiempo elevados. Como resultado, se pierde información debido a la limitación de las encuestas tanto en el espacio como en el tiempo (falta de penetración en espacios marginales, falta de datos en periodo nocturno, etc.).

En este panorama las TIC surgen como fuentes de datos dinámicas y de alta resolución espacio-temporal. El auge de estas nuevas fuentes de datos abre un nuevo abanico de posibilidades en la investigación en la movilidad metropolitana. Esta tesis ha analizado las características, ventajas e inconvenientes de las nuevas fuentes de datos basadas en las TIC como método alternativo para disponer de datos de los que obtener información sobre la movilidad de los ciudadanos y que permitan complementar la información proporcionada por las fuentes tradicionales. Una gran ventaja es la gran riqueza espacio-temporal que presentan estos datos, permitiendo ampliar las temáticas de estudio, obtener datos de cualquier lugar del área de estudio en cualquier momento del día, y realizar análisis comparados respecto a otros espacios u otros momentos temporales.

Además, gracias a las características del *Big Data*, es posible obtener estos datos de forma masiva en tiempo casi real, por lo que se puede obtener una gran cantidad de datos, y se puede actualizar constantemente la información para tener de forma inmediata los patrones de movilidad que vive una ciudad. Otra ventaja importante es la posibilidad de contar, en algunos casos, con estos datos a un coste bajo o gratuito.

Sin embargo, hay que tener en cuenta una serie de limitaciones. Los datos de las TIC son datos desestructurados que no se crean con el objetivo de estudiar la movilidad urbana.

Un desafío en el uso de los datos masivos consiste en el desarrollo de una metodología adecuada para transformar los datos en información adecuada para el desarrollo de planes de movilidad. Hay que asegurar en todo momento la seguridad y privacidad de los datos. Además, estos datos suelen estar sesgados, ya que suelen ser usados principalmente por población en edad joven o con un nivel alto o medio de renta.

Por tanto, la respuesta a la pregunta de investigación es que las nuevas fuentes de datos son adecuadas para el estudio de la movilidad urbana debido a su alto detalle espacio-temporal, y la posibilidad de obtener en muchos casos un amplio volumen de datos a bajo coste. Pese a que estos datos cuentan con algunas limitaciones, se pueden mitigar mediante técnicas como la agregación o el enriquecimiento de datos con información de otras fuentes de datos basadas en las TIC, o con fuentes de datos tradicionales como las encuestas o los censos de información.

b) ¿Qué aporta cada una de las nuevas fuentes de datos al estudio de la movilidad urbana?

La tesis doctoral ha empleado la clasificación de (Moro, 2016) para clasificar las distintas fuentes de datos basadas en las TIC respecto a su valor semántico y la frecuencia a la que se generan los datos. Las tarjetas de transporte y bancarias son las fuentes de datos con mayor riqueza semántica, ya que cuentan con rica información socioeconómica de sus usuarios, pero esta información no suele estar disponible, y además las tarjetas presentan un uso menor que las dos fuentes de datos más utilizadas por los investigadores: los datos de telefonía móvil y las redes sociales.

Uno de los factores que ha contribuido a la proliferación del *Big Data* es la tecnología móvil. Hoy casi toda la población cuenta con un *smartphone* con acceso a internet, lo que permite subir datos de todo tipo y georreferenciarlos gracias a la arquitectura GPS. Los datos de telefonía móvil son los más utilizados en estudios de movilidad urbana gracias a su enorme volumen de datos y su alto detalle temporal. Sin embargo, el detalle espacial de estos datos de telefonía es menor, debido a que el punto registrado corresponde con la antena de telefonía móvil que recoge la llamada. Además, estos datos son poco accesibles, haciendo que el investigador dependa de que las empresas le suministren los datos.

La evolución de internet y las plataformas Web 3.0 ha conllevado al crecimiento de las redes sociales, programas de interacción, abiertos y accesibles, donde se comparte la información que se sube a la red. Muchas redes sociales cuentan con la posibilidad de

compartir datos geolocalizados a partir de los dispositivos GPS de los móviles, por lo que cuentan con un alto detalle espacial. Además, las redes sociales suelen poseer un detalle semántico mayor que los datos de telefonía móvil. Sin embargo, la precisión temporal de los datos de redes sociales es menor que en los datos de telefonía móvil, al estar basada en el momento en el que un usuario publica un mensaje y ser, por tanto, su uso menor.

Esta tesis argumenta que las redes sociales son adecuadas para el estudio de la movilidad urbana, debido a su alto detalle espacial y riqueza semántica, y aunque su detalle temporal es menor, que por ejemplo los datos de telefonía, es lo suficiente para realizar estudios de movilidad. En concreto, esta tesis defiende que la red social *Twitter*, una de las más extendidas en los países occidentales, es especialmente adecuada debido a su carácter semiestructurado, el fácil tratamiento de sus datos en un SIG y su accesibilidad gratuita, mientras que otras redes sociales como *Facebook* cuentan con datos más desestructurados, y con un menor nivel de accesibilidad.

c) ¿Qué herramientas y técnicas pueden ayudar a convertir los datos de redes sociales (como Twitter) en información y conocimiento sobre movilidad?

Como se ha comentado previamente, uno de los principales retos en el uso de datos masivos consiste en el desarrollo de una metodología a seguir para convertir los datos en información. Aunque hay un modelo básico de pasos a seguir (descarga, almacenamiento, preprocesado, análisis y visualización de los datos), no hay métodos o herramientas que sirvan para todas las investigaciones, ya que depende del tipo de fuentes de datos que se usa y del objetivo del estudio para el que se quieren usar los datos. La presente tesis describe una serie de pasos empleados para el tratamiento de datos de *Twitter* y su conversión en datos útiles para estudiar distintos ámbitos de movilidad.

Esta investigación se ha apoyado en el uso de bases de datos *NOSQL* para el almacenado, organización y extracción de los datos que se descargan. En particular, la base de datos *MongoDB* funciona muy bien con la API de descarga de datos de *Twitter*, almacenando los *tweets* en formato *JSON*, y posibilitando la selección y extracción de determinados datos mediante una serie de consultas. El uso de SIG como *ArcGIS Pro* ha permitido transformar los datos *JSON* en entidades de puntos, y facilitar la limpieza, tratamiento, agregación y enriquecimiento de datos, además de posibilitar la incorporación de información nueva mediante la creación de nuevos campos y la unión de diferentes tablas.

Debido al carácter desestructurado de los datos, los pasos de procesado y limpieza han cobrado una importancia vital para poder extraer información confiable y de calidad. Se han propuesto una serie de filtros a seguir para obtener usuarios válidos, con datos suficientes para poder analizar su movilidad individual, y a la vez se han detectado y eliminado las cuentas *bot* o compulsivas. Estos filtros también permiten discernir la movilidad espacial y temporal de los usuarios de la muestra. Después, se ha realizado una ampliación de la muestra mediante la descarga de los 3200 últimos *tweets* de cada usuario considerado válido para los diferentes estudios realizados, con el objetivo de aumentar la precisión espacial y temporal de la huella digital de los usuarios.

Además, se han realizado tareas de enriquecimiento de datos. Así, los datos han sido cruzados con otras fuentes para aumentar la información espacial (demarcación territorial a nivel de municipio, distrito o barrio) y para agregar información de utilidad que ayude a catalogar la actividad del usuario en cada punto (por ejemplo, datos de usos del suelo). A partir de este momento, la metodología utilizada para los procesos de análisis y visualización de los datos difiere según el caso de estudio. Los SIG ponen a su disposición una serie de herramientas geoestadísticas y de geovisualización que permiten el análisis (análisis exploratorio de los datos, análisis de mínimos cuadrados, análisis de clústeres espaciales, Regresión Geográficamente Ponderada, etc.) y cartografía (mapas de residuos, matrices OD, caminos espacio-temporales, cartografía animada, etc.) de los datos de *Twitter*.

d) ¿Pueden los datos de Twitter ser usados para obtener matrices de viajes en espacios metropolitanos?

Uno de los ámbitos en el que el uso de las nuevas fuentes de datos está teniendo más incidencia es en el análisis de flujos de movilidad, en especial en la obtención de matrices de viajes OD. En esta tesis se ha trabajado con datos de *Twitter* para obtener, visualizar y validar matrices de viajes. Gracias al enriquecimiento de los *tweets* a partir de datos de usos del suelo, fue posible mejorar en la definición de los lugares de origen y destino al seleccionar los mensajes publicados en parcelas con uso del suelo residencial o con actividad orientada al trabajo.

Los resultados obtenidos han permitido identificar las principales zonas de generación y atracción de viajes y la intensidad de los flujos entre las unidades espaciales que conforman el Área Metropolitana de Madrid. Se ha obtenido una predominancia de los flujos centrípetos con origen en los municipios colindantes a Madrid o en los distritos

periféricos de la capital, y con destino en los distritos que conforman la Almendra Central. En menor grado también se han observado flujos entre las grandes ciudades del sur metropolitano, o entre los municipios del Corredor del Henares. Agregando los flujos de viajes por grandes zonas metropolitanas, ha sido posible visualizar de forma más sencilla las relaciones comentadas.

Para la verificación de los resultados obtenidos, las matrices se contrastaron con datos del Consorcio de Transportes de Madrid, utilizando la Encuesta Domiciliaria de Movilidad realizada en el año 2018. A partir de esta validación se ha podido comprobar que los resultados son buenos cuando se trabaja a una escala de municipios y distritos, y mejoran cuando se agregan a nivel de grandes zonas metropolitanas. Los datos de la EDM han servido también para contrastar diferentes fuentes empleadas en la expansión de las matrices, siendo mejor la matriz obtenida mediante datos de población residente sobre los orígenes de los viajes.

A pesar de que se ha podido observar que los datos de *Twitter* son válidos para el diseño de matrices de viajes, se han detectado algunos problemas en los resultados. Estos problemas aparecen relacionados con la casuística especial de los espacios centrales, como es el caso del distrito Centro en Madrid y en menor medida de algunos distritos colindantes de la Almendra Central, ya que son zonas con alta actividad turística o nocturna, y con un uso muy alto de *Twitter*, en muchos casos ligado a estas actividades. Como consecuencia, la matriz de viajes presenta una sobreestimación de usuarios en los flujos que tienen la Almendra Central como destino.

e) ¿Es posible utilizar los datos de Twitter para visualizar la movilidad metropolitana mediante caminos espacio-temporales?

La evolución en los últimos años de los SIG como programa de análisis y computación de datos espaciales, junto con la aparición de las TIC, han propiciado un renacer en la Geografía del Tiempo. Una de las herramientas más empleadas en este campo es el camino espacio-temporal diseñado por Hägerstraand. Este tipo de visualización estaba limitado tradicionalmente por la disponibilidad de datos. Sin embargo, los datos procedentes de telefonía móvil o de redes sociales geolocalizadas, permiten actualmente el procesado de caminos espacio-temporales con una alta resolución espacial y temporal. Esta tesis ha buscado indagar en las oportunidades de los datos de *Twitter* y el uso de caminos espacio-temporales como herramienta de visualización de la movilidad individual de usuarios en distintos tipos de espacios urbanos.

La cartografía realizada tanto en 2D (para mostrar las relaciones espaciales) como en 3D (para mostrar los patrones en el tiempo) ha mostrado la utilidad de los datos *Twitter* para la construcción de caminos espacio-temporales gracias a la alta resolución espacial de los *tweets* en forma de coordenadas xy, la información temporal proporcionada por su registro de fecha completo, y la fácil implementación en un SIG en forma de entidades de puntos. Además, estos datos pueden enriquecerse con datos complementarios, como los usos de suelo del Catastro, para poder visualizar la actividad principal de los usuarios en las distintas zonas de estudio a determinadas horas.

En la tesis se han tomado cuatro espacios como áreas de referencia, y se ha analizado la movilidad asociada a ellos a partir del uso de caminos espacio-temporales. De esta forma se ha podido visualizar la zona residencial de Puente de Vallecas como un área que genera viajes, mientras que las otras tres zonas de estudio son atractores de viajes, con diferencias en el tiempo en el que se producen estas atracciones y en las distancias de las zonas desde donde viajan los usuarios. En la zona de oficinas de Nuevos Ministerios, y la zona de Ciudad Universitaria, los usuarios de *Twitter* pueden llegar de espacios lejanos, pero mientras las llegadas a Ciudad Universitaria están más concentradas en el tiempo, las llegadas a Nuevos Ministerios están más diversificadas. Mientras, la zona del Parque de Retiro también atrae viajes, pero de usuarios que vienen de espacios mucho más cercanos y en mayor medida en las horas de la tarde.

f) ¿Se puede estudiar con los datos de Twitter la movilidad de población vinculada a espacios concretos de la ciudad?

El estudio de la movilidad universitaria es un tema de creciente interés debido a las características particulares que suscitan los campus universitarios como espacios de atracción de viajes. Esta tesis ha explorado el uso de *Twitter* para el estudio de la movilidad de la población universitaria, aprovechando el alto consumo de esta red social por parte del sector universitario, y la facilidad de seleccionar *tweets* ubicados en campus universitarios. A partir de la muestra obtenida se ha podido observar una buena correlación entre los usuarios cuyo campus y residencia han sido estimados, y los datos oficiales obtenidos a partir del Ministerio de Educación, Cultura y Deporte.

Los resultados obtenidos han permitido identificar la universidad y campus de los usuarios encontrados en *Twitter*. También han dado una aproximación acerca de los municipios y distritos donde reside la población universitaria, aportación importante ya que no hay fuentes oficiales accesibles al respecto. Se ha observado que los usuarios de

la muestra suelen habitar en los distritos centrales del municipio de Madrid (lugares con servicios orientados a la población joven), o en distritos o municipios cercanos al campus al que viajan. Por tanto, la proximidad es un factor importante tanto en la elección del campus como también en la elección del lugar de residencia. También se han cartografiado las áreas de influencia de cada universidad. De forma general, se ha observado que las universidades generan áreas de influencia en torno a los distritos o municipios cercanos a sus principales campus universitarios. Además de la cercanía al lugar de residencia, se pueden apreciar como factores que definen las áreas de influencia el número de alumnos matriculados en un campus, el tamaño y caracterización socioeconómica del municipio, y la facilidad de acceso a redes de transporte.

Los resultados también han mostrado como los tiempos de viaje desde el municipio de residencia a la universidad son mucho más bajos si el modo de transporte es privado. Además, los tiempos de viaje en transporte público son más son dispares entre los municipios, mientras que los tiempos de viaje en transporte privado son más homogéneos. En general, cuanto más cerca del centro está el campus, son mejores los tiempos en transporte público, al haber acceso a un mayor servicio y frecuencia de transportes. Se ha observado también un mayor número de estudiantes asignados a universidades privadas en las zonas con un nivel alto de renta como los municipios del oeste del área metropolitana. Además, los residentes en estos municipios más ricos tienen que emplear menores tiempos de viaje para desplazarse a los campus universitarios.

Finalmente, se ha comparado la asignación de los usuarios de *Twitter* a las diferentes universidades con modelos clásicos de probabilidad de asignación, como el modelo gravitatorio de *Huff*. Como resultado, se ha observado un buen ajuste entre el número de usuarios de *Twitter* y el modelo gravitacional, aunque los distritos de la Almendra Central y las áreas de influencia de las universidades públicas han presentado una sobreestimación de usuarios de *Twitter*, mientras que los municipios periféricos y las universidades privadas han contado con una subestimación de usuarios.

g) ¿Son los datos de Twitter útiles para analizar el impacto de eventos en el comportamiento espacial de la población?

Los eventos de masas destacan por generar una gran actividad en internet, por lo que redes sociales como *Twitter* pueden convertirse en fuentes valiosas para analizar su

impacto en una ciudad y ayudar a su gestión. La tesis ha trabajado con datos de *Twitter* para obtener información sobre la huella digital que ha tenido la *World Pride* celebrada en el año 2017 sobre la Almendra Central de Madrid, y se han podido comparar los patrones espacio-temporales del evento con la situación observada en una semana de actividad corriente. Esta investigación puede ser de gran interés ya que muestra una metodología para realizar un conteo de visitantes y hallar su lugar de procedencia. Es interesante destacar que los resultados obtenidos son muy similares a los proporcionados por datos de otras fuentes oficiales o de empresas privadas (como es el caso del lugar de procedencia de los turistas).

Los resultados obtenidos han mostrado un destacable aumento de la actividad en *Twitter* durante la *World Pride* respecto a semanas de actividad habitual en la Almendra Central, especialmente durante la noche. Se han hallado picos de actividad que coinciden con los días de algunas de las actividades más importantes del evento (como la manifestación o la clausura del festival). Pero, además, la geolocalización de los datos de *Twitter* ha permitido conocer la impronta espacio-temporal del evento, que en este caso se ha reflejado en la formación de determinados puntos calientes con un alto número de usuarios y un elevado porcentaje de actividad, principalmente en el barrio de Chueca (el principal espacio del festival), pero también en otros puntos de la ciudad, destacando las principales zonas turísticas del casco histórico. Igualmente, se ha apreciado una disminución de la actividad en puntos importantes de la ciudad, pero alejados del centro (cómo Madrid Río o el estadio Santiago Bernabéu). Al analizar las variaciones espacio-temporales se ha visto como, mientras en una semana habitual el principal foco de usuarios se concentra en la estación de trenes de Atocha, durante la semana del evento diversas secciones del casco histórico, y más concretamente del barrio de Chueca, han llegado a superar en número de usuarios a la estación de Atocha, sobre todo en momentos de la tarde-noche.

Una de las innovaciones realizadas en este caso de estudio ha sido el diseño de una metodología para identificar usuarios que han viajado al festival. Para ello, a partir de la descarga de los 3.200 últimos mensajes de cada usuario, se ha analizado si estos usuarios habían publicado *tweets* en Madrid más allá del periodo temporal del evento, y se han identificado sus lugares de procedencia, ya sea identificando la provincia o país con mayor número de mensajes. Se ha observado que el principal lugar de procedencia de los visitantes era la propia Comunidad de Madrid seguida de las provincias más pobladas de

España. Fuera del país, las regiones con un mayor número de visitantes fueron Estados Unidos, América Latina y Europa occidental. Estos resultados han concordado con los resultados obtenidos por empresas privadas como servicios de viajes.

h) ¿Puede la información semántica de los textos de Twitter ser válida para el estudio del transporte urbano?

Entre las nuevas fuentes de datos basadas en el *Big Data*, *Twitter* destaca por su capacidad para proveer de información valiosa sobre las percepciones, opiniones y sentimientos de los usuarios. Esta información puede ser de gran utilidad para estudiar la percepción que los usuarios tienen del transporte público y para identificar problemas potenciales en la red de transporte. En esta tesis, se ha trabajado con datos no geolocalizados de *Twitter* que son respuestas directas a la cuenta de usuario del Metro de Madrid. De esta forma, fue posible obtener una gran cantidad de *tweets* relacionados con el uso de este modo de transporte público en el Área Metropolitana de Madrid. Sin embargo, estos datos no tenían información directa de la localización del *tweet*. Geolocalizar los mensajes en el área de estudio es esencial para poder realizar un análisis espacial de los datos, por lo que, en este caso de estudio, los datos han sido geocodificados a partir del nombre de las estaciones de Metro encontradas en los textos de los usuarios. Un 12,5% de los datos de la muestra fueron geocodificados, lo que todavía han constituido un número grande de *tweets* útiles.

Para extraer opiniones y sentimientos de los usuarios, esta investigación siguió de forma parcial la metodología utilizada por (Saura & Bennett, 2019) para agrupar *tweets* por temas y añadir un valor de sentimiento. Los resultados han mostrado una importante cantidad de *tweets* (principalmente en días laborables y en horario de mañana) con valor negativo mientras que hay pocos *tweets* positivos, confirmando la hipótesis de que los usuarios de *Twitter* tienden a enviar mensajes a las cuentas oficiales de transporte público para quejarse sobre diversos fallos del servicio.

Las estaciones de Metro con mayor cantidad de usuarios con sentimiento negativo en los diferentes temas han sido localizadas en los distritos centrales de Madrid, o en la línea 6 circular que rodea el área central, y que corresponden con intercambiadores entre diferentes líneas de metro o distintos servicios de transporte público. En este estudio se analizaron cuatro temas principales: puntualidad, confort, averías, y sobresaturación. La puntualidad es el problema principal para los usuarios del Metro, aunque se ha hallado un número mayor de registros que expresan quejas relacionadas al confort en las estaciones

del centro de la ciudad, y un mayor número de usuarios reportando problemas de averías en las estaciones localizadas en las zonas periféricas. Por último, se ha desarrollado un modelo de Regresión Geográficamente Ponderada para analizar cómo afectan algunas variables apoyadas en datos oficiales con la distribución espacial de los usuarios de *Twitter* con sentimientos negativos. El mapeado espacial de los coeficientes de las distintas variables ha mostrado que los factores que más han influido en la esta distribución han sido la densidad de puntos de interés (asociada a servicios localizados en los destinos de viajes) y la capacidad intermodal de las estaciones.

5.2. Conclusiones finales

El crecimiento de las ciudades en las últimas décadas ha propiciado nuevos patrones de comportamiento de la movilidad, caracterizados por un mayor número de viajes, más largos y dispersos. La movilidad metropolitana es un campo de investigación bastante dinámico debido a la continua expansión de las redes de transporte y la aparición de planes de actuación urbana que posteriormente se transforman en barrios nuevos. En esta coyuntura, los patrones de movilidad que se desarrollan en la actualidad son diferentes a los comportamientos observados hace unos años. Para poder administrar y gestionar efectivamente las infraestructuras y recursos de una aglomeración, las instituciones requieren de datos constantes y actualizados que se ajusten a los continuos cambios que el área metropolitana sufre en su estructura y en sus redes de transporte. En este panorama, las TIC se antojan como fuentes de datos de gran interés debido a la posibilidad de realizar estudios de movilidad con una cantidad importante de datos, en poco tiempo, a distintas escalas y con un alto nivel de detalle espacio-temporal. Además, en algunos casos se pueden conseguir estos datos de forma gratuita.

El objetivo general de esta tesis doctoral ha sido analizar la movilidad del Área Metropolitana de Madrid en diferentes escenarios, a partir de las nuevas fuentes de datos como *Twitter*, cuya base se sustentan en las TIC y en la participación ciudadana, con el propósito de analizar su validez como fuentes de datos complementarias a las fuentes tradicionales. Se busca responder a la pregunta final de investigación:

i) ¿Hasta qué punto son los datos de Twitter válidos para el estudio de la movilidad en los espacios metropolitanos?

Ya se ha señalado previamente en varios apartados de la tesis como *Twitter* proporciona datos de forma constante y gratuita, lo que facilita disponer de información fácil de actualizar. Aunque las fuentes de datos tradicionales como las EDM tienen información muy valiosa y completa sobre la movilidad de los ciudadanos, el elevado coste de preparación y trabajo de campo que conlleva obtener estos datos impide tener información constante. Por tanto, los datos de *Twitter* se antojan de gran interés para obtener información complementaria a las EDM en los periodos que se dan entre la publicación de las encuestas.

Para validar el valor de *Twitter* como fuentes de datos complementarias a las encuestas, se ha estudiado el potencial de los datos de *Twitter* en sus diferentes dimensiones (espacial, temporal, y semántica). El alto detalle espacio-temporal que poseen estos datos ha permitido analizar la movilidad metropolitana a partir de la visualización de los flujos de viajes de la residencia a los espacios de trabajo o estudios, y de la representación de los caminos de movilidad individual que recorren los usuarios a lo largo del día. La visualización de los patrones generales de movilidad puede ayudar a discernir en que lugares o momentos son necesarias determinadas políticas de actuación para mejorar la movilidad metropolitana. Además, los datos de *Twitter* son fáciles de enriquecer con información externa como los datos de usos del suelo, lo que ayuda a aproximarse a la actividad que realizan los usuarios en diferentes coordenadas espacio-temporales. Esto puede ayudar a discernir los motivos de movilidad de los usuarios a distintas horas.

Los datos de *Twitter* permiten observar los factores que afectan a la movilidad que atraen determinados espacios, como los campus universitarios. Esta información es de gran interés para los organismos públicos o privados, ya que ayuda a la gestión de servicios como horarios de transporte. El detalle espacio-temporal de los datos de *Twitter* ha permitido además analizar el impacto de los eventos sobre una ciudad, gracias a la importancia del uso de redes sociales durante la celebración de estos festivales. Como resultado, los ayuntamientos pueden estimar de antemano el número de visitantes y preparar los servicios de recepción o limpieza con los que asegurar el correcto funcionamiento de la ciudad. Por último, la posibilidad de analizar opiniones sobre los modos de transporte usados a partir de los textos de los *tweets* permite hallar problemas en puntos concretos de una red de transporte y mejorar la oferta de los servicios suministrados.

El contraste con otras fuentes oficiales de datos ha permitido validar la precisión de los datos *Twitter* para estudios de la movilidad urbana. Sin embargo, los datos de *Twitter* presentan algunas limitaciones como la baja muestra de usuarios en algunos casos de estudio. Este problema se debe principalmente al sesgo de la red social y a la poca cantidad de usuarios de *Twitter* que activan la opción de geolocalizar sus *tweets*. Los sesgos son una limitación a tener en cuenta ya que los usuarios de *Twitter* que activan la opción de geolocalización en sus teléfonos móviles suelen ser personas jóvenes, con un nivel socioeconómico alto o medio-alto, y acceso a una mayor cantidad de espacios de actividad. Además, los usuarios de *Twitter* pueden tener diferentes hábitos de escribir mensajes dependiendo de donde viajen (por ejemplo, una mayor predisposición a publicar mensajes en sitios turísticos) (Q. Wang, Phillips, Small, & Sampson, 2018). Precisamente, otro problema a destacar es la sobrerepresentación de la muestra en el distrito Centro de Madrid, concordando con el fuerte carácter turístico, comercial, y de ocio que tiene este distrito (aunque esta sobrerepresentación se puede aprovechar para estudiar aspectos que suceden en el centro de la ciudad como eventos de masas). En cualquier caso, algunos de los problemas detectados pueden ir subsanándose en el tiempo al continuar descargando la información en *streaming*, ya que se incrementa el número de usuarios y localizaciones de la muestra, permitiendo una mayor precisión en trabajos de futuras líneas de investigación.

Por último, esta tesis doctoral puede ayudar a brindar información específica a técnicos y gestores de cara a la contratación de estudios de movilidad basados en datos de *Twitter*. Una consideración práctica a sugerir es que para obtener los mayores beneficios de las nuevas fuentes de datos es necesario modificar la composición de los equipos de planificación y orientarlos a cuadros técnicos multidisciplinarios. Los planificadores del transporte no cuentan con la formación necesaria para trabajar con datos masivos, y los analistas de datos masivos no suelen tener conocimientos en la planificación de transporte. La explosión del *Big Data* hace necesaria la incorporación de expertos en el manejo y análisis de datos masivos y que cada perfil tenga un acercamiento al campo teórico del otro para alcanzar una colaboración eficaz tanto en el tratamiento de datos como en las metodologías de análisis y visualización que ayuden a entender los patrones de movilidad (Gutiérrez-Puebla et al., 2019).

5.3. Futuras líneas de investigación´

El campo del *Big Data* para el estudio de la movilidad urbana todavía está en desarrollo. Esta tesis ha encontrado una serie de limitaciones y retos a superar en futuras investigaciones. Los sesgos de las nuevas fuentes de datos como *Twitter* son limitaciones a mitigar y superar, principalmente su uso predominante por una población estimada en ciudadanos de entre 20 y 39 años, y la menor disponibilidad de datos en las zonas con un bajo nivel de renta (Rashidi et al., 2017). Una futura línea de trabajo en este sentido podría ser tratar de emplear otras fuentes basadas en las TIC como los datos de telefonía o las tarjetas de transporte público, y realizar análisis usando datos conjuntos de esas fuentes.

A la hora de estudiar la movilidad urbana a partir de los flujos de viajes residencia-trabajo, una actualización de los datos del uso del suelo del catastro y una comparación de los resultados en base a los cambios de los usos del suelo pueden contribuir a una mayor precisión de los resultados en próximos trabajos. A su vez, se pueden utilizar diferentes intervalos horarios a los empleados en esta tesis para determinar actividades de residencia y de estudio o trabajo. El aumento del volumen de datos permite actualizar las matrices con un número mayor de *tweets*. A la vez, disponer de muestras mucho mayores de datos permite proponer nuevas investigaciones, como la construcción de matrices de viajes según franjas horarias, grupos sociales, o eventos.

En cuanto a la visualización de caminos espacio-temporales individuales, las actividades diarias de la población se realizan en unas pocas localizaciones habituales (casa, trabajo, etc.), pero ocasionalmente pueden realizarse en localizaciones alternativas. En ocasiones, un individuo puede realizar una actividad diferente, o puede ir a un mismo sitio siguiendo una ruta alternativa (Q. Huang & Wong, 2015). El uso de filtros en función del número de días en los que un usuario ha estado en unas coordenadas espacio-temporales concretas puede ayudar a eliminar puntos aleatorios, mostrando los puntos adecuados para realizar un camino espacio-temporal preciso y concurrente. Un aumento del periodo temporal de la muestra permite obtener un mayor número de usuarios y una mayor cantidad de localizaciones con las que se pueden construir caminos espacio-temporales más precisos. Sin embargo, el aumento de la muestra también conlleva un mayor riesgo de obtener puntos aleatorios. Aumentar el número de días máximos en los que un usuario haya *twitteado* en un lugar a una hora puede disminuir la probabilidad de tener puntos aleatorios. Otras futuras líneas de investigación son la comparación de caminos espacio-

temporales pertenecientes a diferentes colectivos sociales, o la visualización de caminos espacio-temporales durante eventos.

Entre las limitaciones encontradas en el análisis de la movilidad universitaria, están la baja cantidad de *tweets* que se han encontrado en las parcelas correspondientes a facultades, el sesgo de la red social (aprovechado para obtener potenciales alumnos) que sigue dificultando la obtención de otros grupos como profesores o trabajadores, o la imposibilidad de tener datos acerca del reparto modal de los viajeros. Algunas futuras líneas de investigación pueden ser la búsqueda e incorporación de datos que permita obtener el reparto modal de los usuarios, o la comparación de la movilidad de los estudiantes con la del cuerpo docente.

Respecto al análisis del impacto de eventos de masas como la *World Pride 2017*, disponer de muestras mucho mayores de datos puede permitir investigaciones en otros campos como la distribución de visitantes por sexo o edad, el estudio de sentimientos relacionados con el evento a partir del análisis semántico de los textos, o la evaluación de los *hashtags* más comunes durante el evento. Otras futuras líneas de investigación consisten en analizar eventos de otro tipo (por ejemplo, la final de la Liga de Campeones de fútbol 2019/20 celebrada en Madrid el sábado 30 de mayo de 2019) y comparar el impacto causado por eventos con diferentes temáticas.

Finalmente, en el tratamiento de los textos de los *tweets* para la extracción de percepciones sobre el transporte público, aunque la muestra de datos no geolocalizados es mayor y contiene menor ruido que una muestra geolocalizada, la precisión espacial es menor, y en el proceso de limpieza se pierden los datos que no cuentan con información semántica para su geolocalización. Otro problema se halla en la precisión de las técnicas de minería de texto para extraer temas y sentimientos. Aunque las técnicas como la geocodificación, el modelo LDA, o los algoritmos de extracción de sentimientos son útiles para extraer datos, son difíciles de implementar debido a la naturaleza específica y descontextualizada de los textos de los *tweets*, por lo que métodos complementarios como los diccionarios de abreviaturas son necesarios. Algunas futuras líneas de investigación son el análisis comparativo de la red de Metro con otros servicios de transporte público como el Cercanías, el uso combinado de datos no geolocalizados con datos geolocalizados, la identificación de temas más específicos, o el uso de series temporales más largas para explorar patrones anuales e influencias de las situaciones excepcionales y los eventos sobre la red de metro.

5. CONCLUSIONS

The final chapter of this doctoral thesis shows the conclusions obtained from the investigation in the form of answers to the research questions formulated in section 1.2, together with a series of overall conclusions related to the primary objective of the thesis. The chapter concludes with an introduction of lines of research that have opened up during the investigation, and that can be pursued at the postdoctoral stage.

5.1. Answers to research questions

a) Are ICT-based data sources adequate for the study of urban mobility?

Data from traditional sources contain a wealth of valuable information on urban mobility, but are both costly and time-consuming to prepare. This results in information being lost due to the spatial and temporal limitations of such surveys (lack of penetration in marginal areas, lack of night time data, etc.).

In this context, ICTs are a source of dynamic, high space time resolution data. The rise in these new data sources opens a new range of research possibilities into metropolitan mobility. This thesis has analyzed the characteristics, advantages and disadvantages of new ICT-based data as an alternative source of information on urban mobility and as a complement to the information provided by traditional sources. One of the major advantages of these data are their high space-time resolution, which enables to extend the scope of the study, obtain data from any study area at any time of the day and compare them with data obtained at other time points or at other locations.

In addition, thanks to the characteristics of *Big Data*, it is possible to download these data on a massive scale almost in real time, giving access to vast amounts of information that can be constantly updated to obtain city-wide mobility patterns. Another important advantage is that these data are in some cases low cost, or even free of cost.

However, they have a number of limitations. ICT data are de-structured, and are not generated for the purpose of studying urban mobility. One of the challenges in using *Big Data* is to develop an adequate methodology by which the obtained data can be transformed in useful information for the development of mobility plans while guaranteeing data security and privacy. In addition, these data are usually biased, since they are mainly generated by young age population or high- or middle-income individuals.

The answer to the research question, therefore, is that the new data sources are suitable for the study of urban mobility due to their high space-time resolution and the possibility of obtaining, in many cases, a large volume of low-cost data. Although these data have some limitations, these can be mitigated by techniques such as data aggregation or enrichment with information from other ICT-based data sources or with traditional data sources such as surveys or censuses.

b) What does each new data source contribute to the study of urban mobility?

This doctoral thesis has used the method developed by (Moro, 2016) to classify different ICT-based data sources with respect to their semantic value and the frequency at which the data is generated. Travel cards and credit cards are the best source of semantic data because they provide a wealth of social and economic information on their users. However, this information is rarely available, and other data sources, namely, mobile phones and social media, are more widely used in daily life.

Mobile technology is one of the factors that has contributed to the proliferation of *Big Data*. Nowadays nearly everyone has a smartphone with internet access, allowing the upload and georeference of data thanks to the GPS architecture. Mobile phone data is the most widely used in urban mobility studies thanks to its enormous volume of data and its high temporal detail. However, their spatial detail is lower, because the registered point corresponds to the mobile phone antenna that picks up the call. Furthermore, these data are not very accessible, making the researchers dependent of companies supplying the data.

The evolution of the internet and Web 3.0 platforms has led a proliferation of social networks - open, accessible, interactive programs where any uploaded information is shared on the internet. Many social networks can share geotagged data using the GPS devices in mobile phones, so their data has high spatial detail. Also, social networks usually provide more semantic information than mobile phone data. However, social network data has lower temporal detail than mobile phone data, since they are based on the time when a user publishes a message, and count therefore with lower use.

This thesis argues that social networks can be used to study urban mobility due to their high spatial resolution and the wealth of semantic information, and although its temporal resolution is lower than telephony data, it is enough to carry out mobility studies. Specifically, this thesis suggests that *Twitter*, one of the most widespread social networks

in western countries, is particularly suitable for this purpose because its data is semi-structured, can easily be processed in a GIS, and are freely accessibility, while data from other social networks such as *Facebook* are less structured and less accessible.

c) What tools and techniques can be used to convert data from social networks (like Twitter) into information and knowledge about mobility?

As previously mentioned, one of the main challenges in the use of *Big Data* is the development of a methodology of converting data into information. Although there is a basic procedure (download, storage, pre-processing, analysis and visualization of the data), there are no methods or tools suitable for all the investigations, since the procedure will depend on the type of data sources used and the objective of the study. This thesis describes a series of steps used to process data from *Twitter* and convert them into useful data that can be used to study different areas of mobility.

This investigation has used a *NOSQL* database to store, organize and extract the downloaded data. Specifically, the *MongoDB* database works well with the *Twitter* data download API. It stores the *tweets* in *JSON* format and allows certain data to be selected and extracted using a series of queries. Using a GIS such as *ArcGIS Pro*, it was able to transform the *JSON* data into point features, clean, process, aggregate and enrich the data, and also incorporate new information by creating new fields and combining different tables.

Because of the unstructured nature of the data, the processing and cleaning steps were vitally important to extract reliable and quality information. A series of filters has been used to both detect and eliminate *bots* and obtain valid users with sufficient data to analyze their individual mobility. These filters can also discern the spatial and temporal mobility of the sample users. Subsequently, an expansion of the sample was carried out by downloading the last 3200 *tweets* of each user considered valid for the different studies of the thesis, with the aim of increasing the spatial and temporal precision of the users' fingerprint.

In addition, the data has been enriched by cross-referencing them with other sources in order to increase the volume of spatial information (territorial demarcation by town, district or neighborhood) and to aggregate useful information that would help catalogue the user's activity at each point (for example, land registry data). After this, a different methodology of data analysis and visualization was used in each case study. GIS own a

series of geostatistical and geo-visualization tools that allow the analysis (exploratory data analysis, least-squares analysis, spatial cluster analysis, geographically weighted regression, etc.) and mapping (residues maps, OD matrices, space-time trajectories, animated cartography, etc.) the *Twitter* data.

d) Can Twitter data be used to obtain travel matrices in metropolitan areas?

One of the fields that has benefitted most from the use of new data sources is mobility flows analysis, particularly in obtaining OD travel matrices. In this thesis *Twitter* data has been used to obtain, visualize and validate travel matrices. Cross-referencing *tweets* with land registry data enabled to improve the accuracy of origin and destination places by selecting messages originating from residential zones or from work-oriented activity places.

The results obtained have allowed to identify the main generation and attraction zones and the intensity of travel flows between the spatial units that make up the Madrid Metropolitan Area. It has been observed a predominance of centripetal flows originating in the municipalities adjacent to Madrid or in the peripheral districts of the capital, and travelling to the districts located in the center of Madrid, known as the *Almendra Central*. It has been also observed, albeit to a lesser extent, travel flows between the major southern cities of Madrid Metropolitan Area and between towns located in the *Corredor del Henares.* Aggregating travel flows by major metropolitan areas has simplified the visualization of the foregoing correlations.

To verify the results obtained, the matrices were compared with data from the Madrid Transport Consortium using the Household Mobility Survey carried out in 2018. This validation showed that the results were good at the district and municipality scale, and were even better at the scale of large metropolitan areas. The Household Mobility Survey data were also useful for comparing the different sources used to expand the matrices, and showed that the best matrix was obtained using resident population data on the travel origins.

Although *Twitter* data have been shown to be valid for designing travel matrices, some problems have been detected in the results. These problems are related with the special casuistic of the central areas, like the district *Centro,* with high-intensity tourism and night-time activity, and therefore high-intensity *Twitter* usage associated with these

activities. As a result, the travel matrix overestimates the number of users travelling to the *Almendra Central*.

e) Can Twitter data be used to visualize metropolitan mobility using space-time paths?

Recent advances in the use of GIS for analyzing and mapping spatial data, together with the appearance of ICT, have rekindled interest in Time Geography. One of the most widely tools used in this field is Hägerstrand's space-time path. This type of visualization has traditionally been limited by data availability. However, data from mobile phones or geotagged social networks now allow space-time trajectories to be calculated to a high level of space-time resolution. This thesis has investigated the opportunities of using *Twitter* data and space-time paths to visualize the individual mobility of users in different types of urban spaces.

Both 2D (to show spatial relationships) and 3D (to show patterns over time) cartography has shown that *Twitter* data can be used to construct space-time trajectories, thanks to the high spatial resolution of *tweets* in the form of xy coordinates, the time data provided by their complete date stamp, and the ease with which they can be incorporated in a GIS in the form of point features. These data can also be enhanced with complementary data, such as land use data from the Land Registry, in order to visualize the main activity of *Twitter* users in the different study areas at a particular time.

In this thesis, four reference areas have been chosen, and the mobility associated with these zones has been analyzed using time-space paths. This allowed to visualize the *Puente de Vallecas* residential district as a trip generator while the other three study areas as trip attractors, with differences in the time in which these attractions happen and in the distance from the areas from which users travel. *Twitter* users may travel from distant areas to the *Nuevos Ministerios* office district and the *Ciudad Universitaria* university district, but arrivals at *Ciudad Universitaria* are more concentrated in time while arrivals at *Nuevos Ministerios* are more diversified. The *Parque de Retiro* district is also a trip attractor, but attracts users from nearby districts and to a greater extent in the afternoon hours.

f) Can Twitter be used to study population mobility linked to specific areas of the city?

The study of university mobility is a topic of growing interest due to the particular characteristics of university campuses as trip attractors. This thesis has explored the use of *Twitter* to study the mobility of the university population, taking advantage of the high

consumption of this social network by the university sector and the ease of selecting *tweets* located on university campuses. The sample data obtained showed a good correlation between the users whose campus and residence have been estimated and the official data obtained from the Ministry of Education, Culture and Sports.

The results obtained allowed to identify the university and campus of the identified *Twitter* users. The data also indicate the municipalities and districts where the university population lives, a particularly valuable contribution given the lack of available official data in this respect. It has been shown that the sample users usually live around the city center (which offers services oriented to young population), or in districts or municipalities near the campus to which they travel. Therefore, proximity is an important factor in both the choice of campus and the choice of residence. The areas of influence of each university have also been mapped. In general, it has been shown that universities generate areas of influence around the districts or municipalities near their main university campuses. Other factors that determine the university areas of influence, apart from proximity to the place of residence, are the number of students enrolled in a particular campus, the size and economic status of the municipality, and the ease of access to transportation networks.

The results have also shown that travel times from the municipality of residence to the university are far shorter when the transport mode is private. Also, travel times on public transport vary between municipalities, while travel times using private transport are more homogeneous. Generally speaking, the closer the campus is to the metropolitan center, the shorter the travel times on public transport are due to the greater availability and frequency of transport options. It has also been observed that high-income areas, such as the municipalities to the west of the metropolitan area, had more students enrolled in private universities, and that residents in these suburbs took less time to travel to their university campuses.

Finally, classic probability models have been used, such as Huff's gravity model, to compare the allocation of *Twitter* users to the different universities. As result, it has been found a good fit between the number of *Twitter* users and the gravity model, although districts in the *Almendra Central* and the areas of influence of public universities have overestimated the number of *Twitter* users, while users have been underestimated in the peripherical municipalities and in private universities.

g) Are Twitter data useful for analyzing the impact of events on the spatial behavior of the population?

Mass events generate considerable activity on the internet, so social networks like *Twitter* can become valuable tools for analyzing and managing their impact in a city and help in their administration. This thesis has collected *Twitter* data to determine the digital fingerprint of the *World Pride* festival held in 2017 on Madrid's *Almendra Central*, and has compared the event's space-time patterns with the activity observed in a normal week. This investigation can be highly useful, since it shows a methodology of counting visitors and determining their place of origin developed in this study. It is interesting to note that the results are very similar to those obtained from other data provided by government agencies or private corporations (as is the case of the place of origin of tourists).

The results show a notable increase in *Twitter* activity during the *World Pride* festival compared to weeks of normal activity in the *Almendra Central*, particularly at night. Activity peaks have been found to coincide with the days on which some of the most important activities were held (such as the parade or the closing ceremony). In addition, the geotagged *Twitter* data has revealed the space-time imprint of the event, which in this case has been shown by the formation of certain hot spots with a high number of users and a high percentage of activity, mainly in the *Chueca* district (the main forum of the festival), but also in other parts of the city, particularly the main tourist areas of the historic center. This was accompanied by a decrease in activity in key city locations located at some distance from the center (such as the *Madrid Río* park or the *Santiago Bernabéu* stadium). An analysis of space-time variations has shown that while in a typical week most users were concentrated in the *Atocha* train station, during the week of the event the number of users in various parts of the historic center, specifically the *Chueca* district, was higher than in the *Atocha* station, particularly in the evening.

One of the innovations introduced in this case study has been the development of a methodology to identify users who travelled to the festival. Downloading the last 3200 messages from each user has allowed to discern whether they had published *tweets* in Madrid outside the time period of the event, and has enabled to determine their place of origin by identifying either the province or country with the highest number of messages. The results showed that most visitors came from the Community of Madrid, followed by the most populated provinces of Spain. Outside Spain, the regions with the highest number of visitors were the United States, Latin America and Western Europe. These

results have been consistent with data obtained from private companies, such as travel services.

h) Is the semantic information in Twitter texts valid for the study of urban transport?

Twitter stands out from other new *Big Data* sources for its capacity to provide valuable information on user perceptions, opinions and sentiments. This information can be extremely useful to study their opinion of public transport and to identify potential problems in the transport network. This thesis has used non-geotagged *Twitter* replies to the Madrid Metro user account. This has allowed to obtain a large number of *tweets* related to the use of this mode of public transport in the Madrid Metropolitan Area. However, these data did not have information related with the location of the *tweet*. Geotagging messages in the study area is essential in order to perform data spatial analysis, so in this case study the data has been geocoded using the name of the Metro stations found in the users' texts. Although only 12.5% of data were geocoded, this provided a large number of useful *tweets*.

To extract the users opinions and sentiments, this thesis has adapted the methodology used by (Saura & Bennett, 2019) to group *tweets* by topics and add a sentiment score. The results have shown a significant number of negative *tweets* (mainly on weekdays and in the morning) and few positive *tweets*, confirming the hypothesis that *Twitter* users tend to send messages to official public transport accounts to complain about various service failures.

The Metro stations with the largest number of users expressing negative sentiments on different topics have been located in the central districts of Madrid or on the circular line 6 that surrounds the central area, and correspond to interchange stations between different metro lines or different public transport services. In this thesis, four main topics have been analyzed: punctuality, comfort, breakdowns, and overcrowding. Punctuality is the main problem for Metro users, although a greater number of *tweets* have been found complaining about comfort issues in the city central stations, while stations located in the outskirts owned a greater number of users reporting breakdown problems. Finally, a Geographically Weighted Regression model has been developed to analyze how some variables supported by official data affect the spatial distribution of *Twitter* users expressing negative sentiments. Spatial mapping of the coefficients of the different variables has shown that the factor that most influenced this distribution was the density

of points of interest (associated with services located in travel destination) and the intermodal quality of the metro stations.

5.2. Final conclusions

The growth of cities in recent decades has created new mobility patterns characterized by more trips, longer and more dispersed. Metropolitan mobility is a very dynamic field of research due to the continuous expansion of transport networks and the appearance of urban development plans that subsequently become new neighborhoods. This has given rise to mobility patterns that differ from behaviors observed some years ago. In order to effectively administer and manage an agglomeration infrastructures and resources, institutions require a constant flow of updated data that reflect the ongoing changes in the structure and transport networks operating in the metropolitan area. In this context, ICTs can be an extremely useful source of data because they can be used to rapidly conduct mobility studies at different scales using large amounts of high space-time resolution data. In some cases, these data can even be obtained free of charge.

The general objective of this doctoral thesis has been to analyze mobility patterns in the Madrid Metropolitan Area in different scenarios, using new ICT-based data sources such as *Twitter*, which is driven by public participation, in order to analyze whether they can validly complement traditional data sources. The aim has been to answer the ultimate study question:

i) To what extent are Twitter data valid for the study of mobility in metropolitan areas?

Previously this thesis has described how *Twitter* provides a stream of free data that makes it easier to have information that can be easily updated. Although traditional data sources such as mobility surveys contain a great deal of valuable, comprehensive information on the mobility of citizens, the high cost of preparation and fieldwork that this data entails prevents having constant information. Therefore, *Twitter* data are a useful source of information that can fill in the time gaps between the publication of mobility surveys.

To validate the value of *Twitter* as a data source to complement the mobility surveys, *Twitter* data has been analyzed in different dimensions (space, time, and semantic). The high space-time resolution of these data has facilitated the analysis of metropolitan mobility based on visualizing home-work/study travel flows and representing the individual trajectories of users throughout the day. The visualization of general patterns

of mobility can help determine when or where measures should be taken to improve metropolitan mobility. *Twitter* can also be easily enriched with external information such as land use data, thus giving a clearer picture of the users' activity in different space-time coordinates. This can help explain why users travel at certain times of the day.

Twitter data has allowed to observe the factors affecting certain trip attraction places, such as university campuses. This information is of great interest to public or private organizations, as it helps the management of services like transport schedules. The space-time information obtained from *Twitter* data also has enabled to analyze the impact of events on the city, thanks to social networks being used extensively during such festivals. This can help city councils anticipate the number of visitors to a given event, and prepare the reception or cleaning services needed to run the city effectively. Finally, analyzing user opinions on transport modes expressed in the texts of the *tweets* can pinpoint problems at specific points in a transport network and improve the provided services.

Comparing the results with other official data has allowed to validate the accuracy of *Twitter* data in the study of urban mobility. However, *Twitter* data have some limitations, such as the low sample size in some case studies. This is mainly due to the bias inherent to social networks and the low number of *Twitter* users who geotag their *tweets*. Bias are an important limitation, since *Twitter* users who activate the geolocation option on their mobile phones are usually young people, with a high or medium-high socioeconomic level, and access to a greater number of activity spaces. Additionally, *Twitter* users may have different message-writing habits depending on where they travel (for example, a greater predisposition to post messages on tourist sites) (Q. Wang et al., 2018). Precisely, another problem has been the over-representation of the sample in the district *Centro* of Madrid, which is a focus for tourist, commercial and leisure activity (although this overrepresentation can be used to study certain aspects that happen in the city center, such as mass events). Nonetheless, some of the problems detected can be corrected by continuing to download the data in streaming, since the number of users and locations of the sample increases, allowing greater precision in future research lines.

Lastly, this doctoral thesis can help provide specific information to technicians and managers in the hiring of mobility studies based on *Twitter* data. A practical consideration to take into account to maximize the benefits of these new data sources would be to change the composition of planning teams to multidisciplinary technical panels. Transportation planners are not trained to work with *Big Data*, and *Big Data* analysts are

not usually knowledgeable on transportation planning. The boom in *Big Data* compels organizers to include experts in the management and analysis of this type of information, ensuring that each team member has theoretical knowledge of other members' fields, and thus achieving effective collaboration in data processing and in the analysis and visualization methodologies that can help understand mobility patterns (Gutiérrez-Puebla et al., 2019).

5.3. Future lines of research

The use of *Big Data* in the study of urban mobility is still under development. This thesis has outlined a series of limitations and challenges to overcome in future research. The biases inherent to new data sources like *Twitter*, mainly its predominant use among people aged between 20 and 39 years of age and the less availability of data from low-income areas, are limitations that must be mitigated and overcome (Rashidi et al., 2017). A future line of research in this regard could be to conduct analysis using other ICT sources such as mobile phones or travel cards, and perform analysis using joint data from those sources.

When studying urban mobility based on residence-work travel flows, using updated land use data from the land registry and comparing results based on changes in land use can improve the accuracy of future urban mobility studies based on home-work travel flows. Another approach would be to use different time intervals than those used in this thesis to determine home, and work or study activities. Increasing the volume of data can allow researchers to update their matrices with a greater number of *tweets*, while having access to far greater data samples could lead to new research lines, such as constructing travel matrices according to time intervals, social groups, or events.

With regard to the visualization of individual space-time paths, daily activities usually take place in a few locations (home, work, etc.), but can sometimes be carried out at other locations. An individual may sometimes engage in a different activity, or may travel to the same place following a different route (Q. Huang & Wong, 2015). Using filters based on the number of days that a user has been in a specific space-time coordinate can help remove random points, leaving only the correct points needed for an accurate and consistent calculation of the space-time path. Increasing the time period of the sample can provide a greater number of users and a greater number of locations on which to construct more accurate space-time trajectories. However, increasing the sample also carries a

higher risk of obtaining random points. Increasing the maximum number of days in which a user has *tweeted* in a particular place at a particular time can reduce the number of random points. Future lines of research include comparing space-time trajectories among different social groups, or visualizing space-time trajectories during events.

The limitations found in the analysis of university mobility include the low number of *tweets* located in university faculties parcels, the bias inherent to the social network (used in this case to obtain potential students), which continues to make it difficult to obtain users from other groups such as professors or other university employees, or the impossibility of obtaining data on the modal distribution of travelers. Future lines of research may focus on searching for and using data on the modal distribution of users, or on comparing the mobility of students with that of the teaching staff.

In terms of analyzing the impact of mass events such as *World Pride 2017,* obtaining far larger data samples may allow researchers to investigate other fields, such as the distribution of visitors by sex or age, the semantic analysis of the texts to determine sentiments related to the event, or an analysis of the most common hashtags used during the event. Other future lines of research could include analyzing other events (for example, the final of the 2019/20 Champions League held in Madrid on Saturday, 30 May 2019) and comparing the impact caused by events with different themes.

Finally, although the sample of non-geotagged *tweets* processed to extract opinions on public transport is larger and contains less noise than a geotagged sample, the spatial detail is lower, and in the cleaning process data that do not have semantic information that can be used for geolocation are lost. Another problem concerns the accuracy of data mining techniques used to extract topics and sentiments. Although techniques such as geocoding, the LDA model, or sentiment extraction algorithms are useful for extracting data, they are difficult to implement due to the specific and decontextualized nature of the *tweets*, so complementary methods such as abbreviation dictionaries need to be used. Some future lines of research include a comparative analysis of the Metro network and other forms of public transport such as local train services (*Cercanias*), the use of joint non-geotagged and geotagged data, the identification of more specific topics, or the use of longer time series to explore annual patterns and the effect of exceptional situations and events on the metro network.

PUBLICACIONES, CONGRESOS, Y ESTANCIAS

Los resultados de esta tesis doctoral han sido publicados o están en proceso de publicación en diferentes revistas científicas. A continuación, se enumera una lista de los artículos académicos que conforman esta tesis doctoral.

1. Osorio-Arjona, J. y García-Palomares J.C. (2017). ***Nuevas fuentes y retos para el estudio de la movilidad urbana***. Cuadernos Geográficos, 56(3), 247-267. ISSN: 2340-0129 (DOI: 10.30827/CUADGEO.V56I3.5352) [SJR Impact factor: 0.216, Q3 - Geography, Planning and Development].
2. Osorio-Arjona, J. y García-Palomares J.C. (2019). ***Social media and urban mobility: Using Twitter to calculate home-work travel matrices.*** Cities, 89 (June 2019), 268-280 (DOI: 10.1016/j.cities.2019.03.006) [JCR Impact factor 2018: 3.853, Q1 - Urban Studies].
3. Osorio-Arjona, J. y García-Palomares J.C. (2019). ***Big Data y universidades: análisis de movilidad de los estudiantes universitarios a partir de datos de Twitter.*** Geofocus, 24, 37-57. ISSN: 1578-5157 (DOI: 10.21138/GF.648).
4. Osorio-Arjona, J. (2020). ***Análisis de los patrones espacio-temporales de eventos a partir de datos de Twitter: el caso de la World Pride 2017 en Madrid.*** Estudios Geográficos, 81 (288). ISSN: 0014-1496 (DOI: 10.3989/estgeogr.202047.027). [CiteScore 2018: 0.28, Q4 - Earth-Surface Processes].
5. Osorio-Arjona, J. y García-Palomares J.C. (2020). ***Spatio-Temporal mobility and Twitter: 3D visualization of mobility flows.*** Journal of Maps, 16(1), 153-160 (DOI: 10.1080/17445647.2020.1778549). [JCR Impact factor: 1.836, Q3 - Geography].
6. Osorio-Arjona, J., Horak, J., Svodboda, R., García-Ruíz, Y. (enviado en 2020). ***Social media semantic perceptions on Madrid Metro system: using Twitter data to link complaints to space.*** Sustainable Cities and Society. [JCR Impact factor: 1.1, Q1 – Civil and Structural Engineering].

Además, la investigación realizada a lo largo de esta tesis ha sido presentada para exposición en los siguientes congresos y eventos:

1. Osorio-Arjona, J. (2017). ***Spatial Big Data y movilidad a partir de TICs***. II Taller de Doctorado en Geografía. Madrid, 24 octubre 2017.

2. Osorio-Arjona, J. y García-Palomares J.C. (2017). ***Redes sociales y movilidad urbana: cálculo de matrices origen-destino de viajes a partir de Twitter.*** XLIII Reunión de Estudios Regionales, Sevilla, 15-19 de noviembre de 2017.
3. Osorio-Arjona, J., y García-Palomares J.C. (2018). ***Áreas de influencia de los campus universitarios de la Comunidad de Madrid a partir de datos de Twitter.*** XIII Congreso de Ingeniería del Transporte, Gijón, 6-8 de junio de 2018.
4. Osorio-Arjona, J. y García-Palomares J.C. (2018). ***Mega eventos y redes sociales: análisis del impacto del World Pride 2017 en Madrid a partir de datos de Twitter (póster).*** XVIII Congreso Nacional TIG, Valencia, 20-22 de junio de 2018.
5. Osorio-Arjona, J. y García-Palomares J.C. (2018). ***Redes sociales y movilidad urbana: cálculo de matrices origen-destino a partir de datos de Twitter.*** XVI Congreso de la Población en España, Alicante, 12 - 14 de septiembre de 2018.
6. Osorio-Arjona, J. (2018). ***Spatial Big Data y movilidad en el Área Metropolitana de Madrid a partir de Twitter.*** Conferencias ESRI 2018, Madrid (España), 24-25 de octubre de 2018.
7. Osorio-Arjona, J. (2018). ***Spatial Big Data y movilidad en espacios urbanos a partir de TIC.*** II PhDay Facultad de Geografía e Historia de la Universidad Complutense de Madrid, 20 de noviembre de 2018.
8. Osorio-Arjona, J. (2020). ***Análisis de los patrones espacio-temporales de eventos a partir de datos de Twitter: el caso de la World Pride 2017.*** Coloquio Ciudad Justa y Políticas Urbanas en el Contexto Iberoamericano: Ciudades Globales, Ciudades Turísticas, Universidad Complutense de Madrid, 3-5 de marzo de 2020.
9. Horak, J., Orlikova, L., Osorio-Arjona, J., Svoboda, R. (2020). ***Prostorové regresní modelování s příklady.*** GIS Ostrava 2020 – Prostorová data pro Smart City a Smart Region, VŠB - Technical University of Ostrava, 18-20 de marzo de 2020.
10. Osorio-Arjona, J. (2020). ***Twitter y movilidad espacio-temporal: visualización 3D de flujos de movilidad.*** Congreso Campus Foro de Ingeniería del Transporte, 24-26 de junio de 2020.

Por último, una parte de esta investigación (correspondiente al capítulo 4.5 de la tesis doctoral, y al artículo científico número 6 enumerado en este apartado) ha sido realizada en la ***VŠB - Technical University of Ostrava*** (República Checa) en un periodo de tres meses, desde el 9 de septiembre la 8 de diciembre del año 2019.

REFERENCIAS BIBLIOGRÁFICAS

Abdelhaq, H., Sengstock, C., & Gertz, M. (2013). EvenTweet. *Proceedings of the VLDB Endowment*, *6*(12), 1326–1329. https://doi.org/10.14778/2536274.2536307

Aghaei, S., Nematbakhsh, M. A., & Khosravi Farsani, H. (2012). Evolution of the World Wide Web : From Web 1.0 to Web 4.0. *International Journal of Web & Semantic Technology*, *3*(1), 1–10. https://doi.org/10.5121/ijwest.2012.3101

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, *17*(1), 3–27. https://doi.org/10.1080/10630731003597306

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, *41*(6), 855–864. https://doi.org/10.1177/0165551515602847

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, *58*, 240–250. https://doi.org/10.1016/j.trc.2015.02.018

Anselin, L. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Antoine, É., Jatowt, A., Wakamiya, S., Kawai, Y., & Akiyama, T. (2015). Portraying Collective Spatial Attention in Twitter. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 39–48. https://doi.org/10.1145/2783258.2783418

Ascher, F. (2004). *Los nuevos principios del urbanismo*. Madrid: Alianza Editorial.

Ash, J., Kitchin, R., & Leszczynski, A. (2018). Digital turn, digital geographies? *Progress in Human Geography*, *42*(1), 25–43. https://doi.org/10.1177/0309132516664800

Banister, D. (2008). The sustainable mobility paradigm. *Transport Policy*, *15*(2), 73–80. https://doi.org/10.1016/j.tranpol.2007.10.005

Banister, D. (2011). Cities, mobility and climate change. *Journal of Transport Geography*, *19*(6), 1538–1546. https://doi.org/10.1016/j.jtrangeo.2011.03.009

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies*, *15*(6), 380–391. https://doi.org/10.1016/J.TRC.2007.06.003

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography*, *3*(3), 297–302. https://doi.org/10.1177/2043820613514323

Batista e Silva, F., Marín Herrera, M. A., Rosina, K., Ribeiro Barranco, R., Freire, S., & Schiavina, M. (2018). Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. *Tourism Management*, *68*, 101–115. https://doi.org/10.1016/j.tourman.2018.02.020

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, *3*(3), 274–279. https://doi.org/10.1177/2043820613513390

Biever, C. (2010). Twitter mood maps reveal emotional states of America. *New Scientist*, *207*(2771), 14. https://doi.org/10.1016/S0262-4079(10)61833-7

Birkin, M., Harland, K., Malleson, N., Cross, P., & Clarke, M. (2014). An Examination of Personal Mobility Patterns in Space and Time Using Twitter. *International Journal of Agricultural and Environmental Information Systems*, *5*(3), 55–72. https://doi.org/10.4018/ijaeis.2014070104

BITRE. (2014). *New traffic data sources – An overview*. *New Data Sources for Transport Workshop*. Retrieved from https://www.bitre.gov.au/sites/default/files/2019-12/GHD-report-new-technologies-workshop.pdf

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-Located Tweets. Enhancing Mobility Maps and Capturing Cross-Border Movement. *PLOS ONE*, *10*(6), e0129202. https://doi.org/10.1371/journal.pone.0129202

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). *Latent Dirichlet Allocation Michael I. Jordan*. *Journal of Machine Learning Research* (Vol. 3).

Bloem, J., van Doorn, M., Duivestein, S., Excoffier, D., Maas, R., & van Ommeren, E. (2014). *The Fourth Industrial Revolution. Things to Tighten the Link between IT and OT*.

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., & Smoreda, Z. (2015). Passive

Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations. *Transportation Research Procedia*, *11*, 381–398. https://doi.org/10.1016/j.trpro.2015.12.032

Bosque, J. (2015). Neogeografía , Big Data y Tig : Problemas y nuevas posibilidades. *Polígonos*, *27*(2007), 165–173.

Bregman, S. (2012). *TCRP Synthesis 99: Uses of Social Media in Public Transportation*. -. Retrieved from http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_syn_99.pdf

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Caceres, N., Wideberg, J. P., & Benitez, F. G. (2007). Deriving origin–destination data from a mobile phone network. *IET Intelligent Transport Systems*, *1*(1), 15. https://doi.org/10.1049/iet-its:20060020

Caceres, Noelia, Romero, L. M., Benitez, F. G., & Del Castillo, J. M. (2012). Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems*, *13*(3), 1430–1441. https://doi.org/10.1109/TITS.2012.2189006

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2014). A Scalable Framework for Spatiotemporal Analysis of Location-based Social Media Data. *Computers, Environment and Urban Systems*, *51*, 70–82. Retrieved from http://arxiv.org/abs/1409.2826

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*. https://doi.org/10.1016/j.apgeog.2012.01.005

Casas, I., & Delmelle, E. C. (2017). Tweeting about public transit — Gleaning public perceptions from a social media microblog. *Case Studies on Transport Policy*, *5*(4), 634–642. https://doi.org/10.1016/j.cstp.2017.08.004

Chen, B. Y., Li, Q., Wang, D., Shaw, S. L., Lam, W. H. K., Yuan, H., & Fang, Z. (2013). Reliable Space-Time Prisms Under Travel Time Uncertainty. *Annals of the*

Association of American Geographers, *103*(6), 1502–1521. https://doi.org/10.1080/00045608.2013.834236

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, *68*, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

Chen, J., Shaw, S.-L., Yu, H., Lu, F., Chai, Y., & Jia, Q. (2011). Exploratory data analysis of activity diary data: a space–time GIS approach. *Journal of Transport Geography*, *19*(3), 394–404. https://doi.org/10.1016/j.jtrangeo.2010.11.002

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. In *Mobile Networks and Applications* (Vol. 19, pp. 171–209). Kluwer Academic Publishers. https://doi.org/10.1007/s11036-013-0489-0

Cheng, J., Gould, N., Han, L., & Jin, C. (2016). Big Data for Urban Studies: Opportunities and Challenges: A Comparative Perspective. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (pp. 1229–1234). IEEE. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0189

Cheshire, J., & Uberti, O. (2016). *London : the information capital : 100 maps and graphics that will change how you view the city*. London: Penguin Group.

Chin, D., Zappone, A., & Zhao, J. (2016). Analyzing Twitter Sentiment of the 2016 Presidential Candidates. In *Applied Informatics and Technology Innovation Conference (AITIC 2016)*.

Ciuccarelli, P., Lupi, G., & Simeone, L. (2014). *Visualizing the Data City*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-02195-9

Clarke, M. (2013). *Big Data in Transport*. *Institution of Engineering and Technology Sectors Insights*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137378972

Collins, C., Hasan, S., & Ukkusuri, S. V. (2013). A Novel Transit Rider Satisfaction Metric A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured

from Online Social Media Data. *Journal of Public Transportation*, *16*(2), 21–45.

Conover, M. D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., & Flammini, A. (2013). The Geospatial Characteristics of a Social Movement Communication Network. *PLoS ONE*, *8*(3). https://doi.org/10.1371/journal.pone.0055957

Davison, L., Ahern, A., & Hine, J. (2015). Travel, transport and energy implications of university-related student travel: A case study approach. *Transportation Research Part D: Transport and Environment*, *38*, 27–40. https://doi.org/10.1016/j.trd.2015.04.028

De Domenico, M., Lima, A., & Musolesi, M. (2013). Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, *9*(6), 798–807. https://doi.org/10.1016/j.pmcj.2013.07.008

de las Rivas, J. L., Iglesias, F., & Lalana, J. L. (2011). Campus universitario de Valladolid. Integración urbana y movilidad. *Bitácora*, *18*(1), 139–156.

de Smith, M. J., Goodchild, M. F., & Longley, P. A. (2018). *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*. Retrieved from http://www.spatialanalysisonline.com/

Delclòs-Alió, X., & Miralles-Guasch, C. (2017). Suburban travelers pressed for time: Exploring the temporal implications of metropolitan commuting in Barcelona. *Journal of Transport Geography*, *65*, 165–174. https://doi.org/10.1016/j.jtrangeo.2017.10.016

Delmelle, E. M., & Delmelle, E. C. (2012). Exploring spatio-temporal commuting patterns in a university environment. *Transport Policy*, *21*, 1–9. https://doi.org/10.1016/j.tranpol.2011.12.007

Demšar, U., & Virrantaus, K. (2010). Space-time density of trajectories: Exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, *24*(10), 1527–1542. https://doi.org/10.1080/13658816.2010.511223

Dewulf, B., Neutens, T., Vanlommel, M., Logghe, S., De Maeyer, P., Witlox, F., … Van de Weghe, N. (2015). Examining commuting patterns using Floating Car Data and circular statistics: Exploring the use of new methods and visualizations to study travel times. *Journal of Transport Geography*, *48*(December), 41–51.

https://doi.org/10.1016/j.jtrangeo.2015.08.006

El-Diraby, T., Shalaby, A., & Hosseini, M. (2019). Linking social, semantic and sentiment analyses to support modeling transit customers' satisfaction: Towards formal study of opinion dynamics. *Sustainable Cities and Society*, *49*. https://doi.org/10.1016/j.scs.2019.101578

Fang, Z., Shaw, S.-L., Tu, W., Li, Q., & Li, Y. (2012). Spatiotemporal analysis of critical transportation links based on time geographic concepts: a case study of critical bridges in Wuhan, China. *Journal of Transport Geography*, *23*, 44–59. https://doi.org/10.1016/j.jtrangeo.2012.03.018

Farber, S., Neutens, T., Miller, H. J., & Li, X. (2013). The Social Interaction Potential of Metropolitan Regions: A Time-Geographic Measurement Approach Using Joint Accessibility. *Annals of the Association of American Geographers*, *103*(3), 483–504. https://doi.org/10.1080/00045608.2012.689238

Farber, S., O'Kelly, M., Miller, H. J., & Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography*, *49*(December), 26–38. https://doi.org/10.1016/j.jtrangeo.2015.10.009

Feria Toribio, J. (2010). La delimitación y organización espacial de las áreas metropolitanas españolas: una perspectiva desde la movilidad residencia-trabajo. *Ciudad y Territorio: Estudios Territoriales*, (164), 189–210.

Finnis, K. K., & Walton, D. (2007). Field observations of factors influencing walking speeds. *International Conference on Sustainability Engineering and Science*. Retrieved from https://trid.trb.org/view/835669

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing Urban Landscapes Using Geolocated Tweets. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing* (pp. 239–248). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.19

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gao, S., Yang, J., Yan, B., Hu, Y., Janowicz, K., & McKenzie, G. (2014). Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area. *Eight International Conference on Geographic Information Science (GIScience'14)*. Retrieved from http://www.geog.ucsb.edu/~sgao/papers/2014_GIScience_EA_DetectingODTripsUsingGeoTweets.pdf

García-Albertos, P., Picornell, M., Salas-Olmedo, M. H., & Gutiérrez, J. (2019). Exploring the potential of mobile phone records and online route planners for dynamic accessibility analysis. *Transportation Research Part A: Policy and Practice*, *125*, 294–307. https://doi.org/10.1016/j.tra.2018.02.008

García-Palomares, J. C. (2010). Urban sprawl and travel to work: the case of the metropolitan area of Madrid. *Journal of Transport Geography*, *18*(2), 197–213. https://doi.org/10.1016/j.jtrangeo.2009.05.012

García-Palomares, J. C., & Gutiérrez-Puebla, J. (2007). La ciudad dispersa: cambios recientes en los espacios residenciales de la Comunidad de Madrid. *Anales de Geografía*, *27*(1), 45–67. https://doi.org/10.1016/0360-3016(82)90387-X

García-Palomares, J. C., Gutiérrez, J., & Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography*, *63*, 408–417. https://doi.org/10.1016/j.apgeog.2015.08.002

García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gómez, B., Condeço-Melhorado, A., & Gutiérrez, J. (2018). City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities*, *72*, 310–319. https://doi.org/10.1016/j.cities.2017.09.007

Gasparini, A., & Guidicini, P. (1990). *Innovazione tecnologica e nuovo ordine urbano*. Torino: Angeli.

Ghosh, D., & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2). https://doi.org/10.1080/15230406.2013.776210

Goodchild, M. F. (2007). Citizers as sensors: Web 2.0 and the volunteering of

geographic information, 8–10. Retrieved from www.flickr.com

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer*, *66*(4), 568–578. https://doi.org/10.1080/00330124.2014.907699

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*, *3*(3), 255–261. https://doi.org/10.1177/2043820613513121

Griffiths, R., Richardson, A. J., & Lee-Gosselin, M. E. H. (2000). Travel Surveys. *Transportation in the New Millennium*. Retrieved from https://trid.trb.org/view/639425

Gutiérrez-Puebla, J., Benitez, C., Leaño, J. M., García-Palomares, J. C., Condeço-Melhorado, A., Mojica, C., … Romanillos, G. (2019). *Cómo aplicar Big Data en la planificación del transporte urbano. El uso de datos de telefonía móvil en el análisis de la movilidad*. (C. Benitez, Ed.). Banco Interamericano de Desarrollo.

Gutiérrez-Puebla, J., García-Palomares, J. C., & Salas-Olmedo, M. H. (2016). Big (Geo) Data en Ciencias Sociales: Retos y Oportunidades. *Revista de Estudios Andaluces*, *33*(331), 1–23. https://doi.org/10.12795/rea.2016.i33.0

Gutiérrez Gallego, J. A., Ruiz Labrador, E. E., & Rodrigo Muñoz, R. (2016). Diagnóstico de la movilidad en los campus de la universidad de Extremadura. *XVII Congreso Nacional de Tecnologías de Información Geográfica*, 140–154.

Gutiérrez, J., & García-Palomares, J. C. (2007). New spatial patterns of mobility within the metropolitan area of Madrid: Towards more complex and dispersed flow networks. *Journal of Transport Geography*, *15*(1), 18–30. https://doi.org/10.1016/J.JTRANGEO.2006.01.002

Gutiérrez, J., García-Palomares, J. C., Romanillos, G., & Salas-Olmedo, M. H. (2017). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. *Tourism Management*, *62*, 278–291. https://doi.org/10.1016/j.tourman.2017.05.003

Gutiérrez Puebla, J. (2018). Big Data y nuevas geografías: la huella digital de las actividades humanas. *Documents d'Anàlisi Geogràfica*, *64*(2), 195. https://doi.org/10.5565/rev/dag.526

Hägerstraand, T. (1970). What about people in regional science? *Papers in Regional Science*, *24*(1), 7–24. https://doi.org/10.1111/j.1435-5597.1970.tb01464.x

Haghighi, N. N., Liu, X. C., Wei, R., Li, W., & Shao, H. (2018). Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. *Public Transport*, *10*(2), 363–377. https://doi.org/10.1007/s12469-018-0184-4

Hanabusa, H. (2012). Development of Nowcast Traffic Simulation System for Road Traffic in Urban Area. *20th World Congress on ITS*, *10*, 3–10.

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13* (p. 1). New York, New York, USA: ACM Press. https://doi.org/10.1145/2505821.2505823

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-intensive scientific discovery* (Microsoft). Redmond, WA.

Hiltz, S., Pfaff, M., Plotnick, L., Robinson, A., Caragea, C., Squicciarini, A., … Tapia, A. (2014). Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy. In *Proceedings of the 11th International ISCRAM Conference – University Park, Pennsylvania, USA*. Retrieved from http://reliefweb.int/sites/reliefweb.int/files/resources/report_36.pdf

Hosseini, M., El-Diraby, T., & Shalaby, A. (2018). Supporting sustainable system adoption: Socio-semantic analysis of transit rider debates on social media. *Sustainable Cities and Society*, *38*, 123–136. https://doi.org/10.1016/j.scs.2017.12.025

Huang, Q., & Wong, D. W. S. (2015). Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. *Annals of the Association of American Geographers*, *105*(6), 1179–1197. https://doi.org/10.1080/00045608.2015.1081120

Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30*(9), 1873–1898. https://doi.org/10.1080/13658816.2016.1145225

Huang, Y., Li, Y., & Shan, J. (2018). Spatial-Temporal Event Detection from Geo-Tagged Tweets. *ISPRS International Journal of Geo-Information*, *7*(4), 150. https://doi.org/10.3390/ijgi7040150

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Järv, O., Tenkanen, H., & Toivonen, T. (2017). Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science*, *31*(8), 1630–1651. https://doi.org/10.1080/13658816.2017.1287369

Ji, T., Fu, K., Self, N., Lu, C.-T., & Ramakrishnan, N. (2018). Multi-task Learning for Transit Servide Disruption Detection. In *ASONAM 2018 : proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Jin, X., Long, Y., Sun, W., Lu, Y., Yang, X., & Tang, J. (2017). Evaluating cities' vitality and identifying ghost cities in China with emerging geographical data. *Cities*, *63*, 98–109. https://doi.org/10.1016/j.cities.2017.01.002

Juan M. Albertos, Joan Noguera, María D. Pitarch, J. S. (2008). *Los hábitos de movilidad en la Universitat de València (2005-2006): Problemas de acceso a los campus y sostenibilidad*. Valencia: Universitat de València.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding human mobility from Twitter. *PLoS ONE*, *10*(7), 1–16.

https://doi.org/10.1371/journal.pone.0131469

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. In *2013 46th Hawaii International Conference on System Sciences* (pp. 995–1004). IEEE. https://doi.org/10.1109/HICSS.2013.645

Kang, C., Gao, S., Lin, X., Xiao, Y., Yuan, Y., Liu, Y., & Ma, X. (2010). Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *2010 18th International Conference on Geoinformatics, Geoinformatics 2010*. https://doi.org/10.1109/GEOINFORMATICS.2010.5567857

Kemp, S. (2019). The state of digital in April 2019: all the numbers you need to know. Retrieved October 22, 2019, from https://wearesocial.com/blog/2019/04/the-state-of-digital-in-april-2019-all-the-numbers-you-need-to-know

Keskin, M., Çelik, B., Doğru, A. Ö., & Pakdil, M. E. (2014). *A Comparison of Space-Time 2D and 3D Geovisualization*. Retrieved from https://www.researchgate.net/publication/281643999

Kim, K.-S., Kojima, I., & Ogawa, H. (2016). Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science*, *30*(9), 1899–1922. https://doi.org/10.1080/13658816.2016.1146956

Kirilenko, A. P., & Stepchenkova, S. O. (2017). Sochi 2014 Olympics on Twitter: Perspectives of hosts and guests. *Tourism Management*, *63*, 54–65. https://doi.org/10.1016/j.tourman.2017.06.007

Kitchin, R. (2013). Big data and human geography. *Dialogues in Human Geography*, *3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Knott, B., Swart, K., & Visser, S. (2015). The impact of sport mega-events on the quality of life for host city residents: reflections on the 2010 FIFA World Cup. *African Journal of Hospitality, Tourism and Leisure*, *4*(January), 1–16.

Kocich, D. (2017). Multilingual sentiment mapping using Twitter, Open Source tools, and dictionary based machine translation approach. In *GIS Ostrava*.

Kocich, D., & Horák, J. (2016). Twitter as a source of big spatial data. In *16th International Multidisciplinary Scientific GeoConference SGEM2016, Informatics, Geoinformatics and Remote Sensing* (Vol. 1). Stef92 Technology. https://doi.org/10.5593/SGEM2016/B21/S08.116

Kovacs-Gyori, A., Ristea, A., Havas, C., Resch, B., & Cabrera-Barona, P. (2018). #London2012: Towards Citizen-Contributed Urban Planning Through Sentiment Analysis of Twitter Data. *Urban Planning*, *3*(1), 75. https://doi.org/10.17645/up.v3i1.1287

Kovacs-Györi, A., Ristea, A., Kolcsar, R., Resch, B., Crivellari, A., & Blaschke, T. (2018). Beyond Spatial Proximity—Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data. *ISPRS International Journal of Geo-Information*, *7*(9), 378. https://doi.org/10.3390/ijgi7090378

Kulkarni, G., Abellera, L., & Panangadan, A. (2018). Unsupervised classification of online community input to advance transportation services. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 261–267). IEEE. https://doi.org/10.1109/CCWC.2018.8301704

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data. *PLoS ONE*, *9*(6), 1–15. https://doi.org/10.1371/journal.pone.0096180

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Lansley, G., Smith, M. De, Goodchild, M., & Longley, P. (2018). Big Data and Geospatial Analysis. In *Geospatial Analysis 6th Edition* (pp. 547–570). Edinburgh: The Winchelsea Press.

Lathia, N., Smith, C., Froehlich, J., & Capra, L. (2013). Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*, *9*(5), 643–664. https://doi.org/10.1016/j.pmcj.2012.10.007

Lee, J. H., Goao, S., & Goulias, K. G. (2015). Can Twitter data be used to validate

travel demand models? *GEOTRANS Report 2015-5-03*, 1–27.

Lee, J., & Miller, H. J. (2018). Measuring the impacts of new public transit services on space-time accessibility: An analysis of transit system redesign and new bus rapid transit in Columbus, Ohio, USA. *Applied Geography*, *93*(September 2017), 47–63. https://doi.org/10.1016/j.apgeog.2018.02.012

Lee, J. Y., & Kwan, M.-P. (2011). Visualization of socio-spatial isolation based on human activity patterns and social networks in space-time. *Tijdschrift Voor Economische En Sociale Geografie*, *102*(4), 468–485. https://doi.org/10.1111/j.1467-9663.2010.00649.x

Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks - LBSN '10* (p. 1). New York, New York, USA: ACM Press. https://doi.org/10.1145/1867699.1867701

Lenormand, M., Louail, T., Cantú-Ros, O. G., Picornell, M., Herranz, R., Arias, J. M., … Ramasco, J. J. (2015). Influence of sociodemographic characteristics on human mobility. *Scientific Reports*, *5*(1), 10075. https://doi.org/10.1038/srep10075

Lenormand, M., Picornell, M., Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., … Ramasco, J. J. (2014). Cross-Checking Different Sources of Mobility Information. *PLoS ONE*, *9*(8), 1–10. https://doi.org/10.1371/journal.pone.0105184

Leszczynski, A., & Crampton, J. (2016). Introduction: Spatial Big Data and everyday life. *Big Data & Society*, *3*(2), 205395171666136. https://doi.org/10.1177/2053951716661366

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social Media in Tourism and Hospitality: A Literature Review. *Journal of Travel and Tourism Marketing*, *30*(1–2), 3–22. https://doi.org/10.1080/10548408.2013.750919

Lewis, L. (2019). 2019: This Is What Happens In An Internet Minute. Retrieved September 18, 2019, from https://www.allaccess.com/merge/archive/29580/2019-this-is-what-happens-in-an-internet-minute

Li, H., Ji, H., & Zhao, L. (2015). Social Event Extraction: Task, Challenges and Techniques. In *Proceedings of the 2015 IEEE/ACM International Conference on*

Advances in Social Networks Analysis and Mining 2015 - ASONAM '15 (pp. 526–532). New York, New York, USA: ACM Press. https://doi.org/10.1145/2808797.2809413

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., … Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Lim, K. H., Lee, K. E., Kendal, D., Rashidi, L., Naghizade, E., Winter, S., & Vasardani, M. (2018). The Grass is Greener on the Other Side. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (pp. 275–282). New York, New York, USA: ACM Press. https://doi.org/10.1145/3184558.3186337

Liu, H., Ge, Y., Zheng, Q., Lin, R., & Li, H. (2018). Detecting global and local topics via mining twitter data. *Neurocomputing*, *273*, 120–132. https://doi.org/10.1016/j.neucom.2017.07.056

Long, Y., & Shen, Z. (2015). Profiling Underprivileged Residents with Mid-term Public Transit Smartcard Data of Beijing. In *Geospatial Analysis to Support Urban Planning in Beijing* (pp. 169–192). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-19342-7_9

Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, *30*(2), 369–389. https://doi.org/10.1080/13658816.2015.1089441

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, *47*(2), 465–484. https://doi.org/10.1068/a130122p

Louail, T., Lenormand, M., Cantú, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., … Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. *Scientific Reports*, *4*, 1–12. https://doi.org/10.1038/srep05276

Louail, T., Lenormand, M., Picornell, M., García Cantú, O., Herranz, R., Frias-Martinez, E., … Barthelemy, M. (2015). Uncovering the spatial structure of mobility networks. *Nature Communications*, *6*(1), 1–8. https://doi.org/10.1038/ncomms7007

Lucas-García, F., Racero-Moreno, J., Torrecillas, C., & García-Sánchez, J. M. (2016). Análisis de la movilidad en campus universitarios integrados en zonas urbanas. *Dyna (Spain)*, *91*(3), 1–10. https://doi.org/10.6036/7660

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography*, *70*, 11–25. https://doi.org/10.1016/j.apgeog.2016.03.001

Luong, T. T. B., & Houston, D. (2015). Public opinions of light rail service in Los Angeles , an analysis using Twitter data. *IConference 2015 Proceedings*, 2–5.

Manetti, G., Bellucci, M., & Bagnoli, L. (2017). Stakeholder Engagement and Public Information Through Social Media: A Study of Canadian and American Public Transportation Agencies. *The American Review of Public Administration*, *47*(8), 991–1009. https://doi.org/10.1177/0275074016649260

Marine-Roig, E., & Anton Clavé, S. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing and Management*, *4*(3), 162–172. https://doi.org/10.1016/j.jdmm.2015.06.004

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2017). Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities*, *64*, 66–78. https://doi.org/10.1016/j.cities.2017.02.007

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social Media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, *74*, 161–174. https://doi.org/10.1016/j.compenvurbsys.2018.11.001

Martin, D., Jordan, H., & Roderick, P. (2008). Taking the Bus: Incorporating Public Transport Timetable Data into Health Care Accessibility Modelling. *Environment and Planning A: Economy and Space*, *40*(10), 2510–2525. https://doi.org/10.1068/a4024

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data : a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical*

Information Systems, *5*(3), 287–301. https://doi.org/10.1080/02693799108927856

Miller, H. J. (2005). Necessary space - Time conditions for human interaction. *Environment and Planning B: Planning and Design*, *32*(3), 381–401. https://doi.org/10.1068/b31154

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, *50*(1), 181–201. https://doi.org/10.1111/j.1467-9787.2009.00641.x

Miller, H. J. (2017). Time Geography and Space-Time Prism. In *International Encyclopedia of Geography* (pp. 1–19). Oxford, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118786352.wbieg0431

Miller, H. J., & Goodchild, M. F. (2014). Data-driven geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Miller, H. J., Raubal, M., & Jaegal, Y. (2016). *Geospatial Data in a Changing World*. (T. Sarjakoski, M. Y. Santos, & L. T. Sarjakoski, Eds.). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33783-8

Miralles-Guasch, C. (2012). Las encuestas de movilidad y los referentes ambientales de los transportes. *EURE (Santiago)*, *38*(115), 33–45. https://doi.org/10.4067/S0250-71612012000300002

Miralles-Guasch, C., Delclòs, X., & Vich, G. (2015). Nuevas fuentes de información para el análisis de la movilidad cotidiana: de las encuestas de movilidad a las aplicaciones para móviles. *XXIV Congreso de La Asociación de Geógrafos Españoles*, (January), 2055–2063.

Miralles-Guasch, C., & Domene, E. (2010). Sustainable transport challenges in a suburban university: The case of the Autonomous University of Barcelona. *Transport Policy*, *17*(6), 454–463. https://doi.org/10.1016/j.tranpol.2010.04.012

Miralles-Guasch, C., & Martínez, M. (2013). Las fuentes de información sobre movilidad: la visión de los profesionales. Ejemplo de aplicación de metodología DELPHI. *Revista Transporte y Territorio*, (8), 99–116. Retrieved from http://www.rtt.filo.uba.ar/RTT00806100.pdf

Miralles-Guasch, C., Martínez Melo, M., & Marquet Sarda, O. (2014). On user perception of private transport in Barcelona Metropolitan area: an experience in an

academic suburban space. *Journal of Transport Geography*, *36*, 24–31. https://doi.org/10.1016/j.jtrangeo.2014.02.009

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064417

Moravec Giormenti, B., López Dentone, F., Bossio, D., Filgueira, E. M., Gurrera, W., Piccirillo, J. M., … Caprelli, C. (2018). *La movilidad de los estudiantes de la Facultad Regional Avellaneda de la Universidad Tecnológica Nacional*. Retrieved from www.c3t.fra.utn.com.ar

Moro, E. (2016). Ciudades, Movilidad y Social Media. VII Congreso Estatal RITSI.

Moya-Gómez, B., & García-Palomares, J. C. (2015). Working with the daily variation in infrastructure performance on territorial accessibility. The cases of Madrid and Barcelona. *European Transport Research Review*, *7*(2). https://doi.org/10.1007/s12544-015-0168-2

Moya-Gómez, B., Salas-Olmedo, M. H., García-Palomares, J. C., & Gutiérrez, J. (2017). Dynamic Accessibility using Big Data: The Role of the Changing Conditions of Network Congestion and Destination Attractiveness. *Networks and Spatial Economics*, 1–18. https://doi.org/10.1007/s11067-017-9348-z

Moyano, A., Moya-Gómez, B., & Gutiérrez, J. (2018). Access and egress times to high-speed rail stations: a spatiotemporal accessibility analysis. *Journal of Transport Geography*, *73*, 84–93. https://doi.org/10.1016/j.jtrangeo.2018.10.010

Murthy, D. (2018). *Twitter : social communication in the Twitter age*. Cambridge: Polity Press. Retrieved from https://www.worldcat.org/title/twitter-social-communication-in-the-twitter-age/oclc/1028022733

Netto, V. M., Pinheiro, M., Meirelles, J. V., & Leite, H. (2015). Digital footprints in the cityscape: Finding networks of segregation through Big Data. *International Conference on Location-Based Social Media Data*, (March), 1–15.

Neutens, T., Van de Weghe, N., Witlox, F., & De Maeyer, P. (2008). A three-dimensional network-based space - Time prism. *Journal of Geographical Systems*, *10*(1), 89–107. https://doi.org/10.1007/s10109-007-0057-x

OECD. (2015). Big Data and Transport: Understanding and assessing options. *International Transport Forum*. https://doi.org/10.1002/ajh.23643

Ortúzar S., J. de D., & Willumsen, L. G. (2011). *Modelling transport*. Oxford: Wiley-Blackwell. Retrieved from https://www.wiley.com/en-gb/Modelling+Transport%2C+4th+Edition-p-9780470760390

Pajević, F., & Shearmur, R. G. (2017). Catch Me if You Can: Workplace Mobility and Big Data. *Journal of Urban Technology*, *24*(3), 99–115. https://doi.org/10.1080/10630732.2017.1334855

Pallares-Barbera, M., & Masala, E. (2016). When Internet became Geography. Spatial patterns on urban open spaces through the analysis of user-generated data in Barcelona. *Urban Transitions and Economic Spaces View Project Economic Geography and Big Data View Project*. Retrieved from https://www.researchgate.net/publication/301642878

Pazos, M. (2005). El estudio de la movilidad diaria en España: limitaciones en las fuentes y alternativas propuestas. *Eria*, *66*, 85–92.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2014.913794

Pereira, F. C., Rodrigues, F., & Ben-Akiva, M. (2015). Using Data From the Web to Predict Public Transport Arrivals Under Special Events Scenarios. *Journal of Intelligent Transportation Systems*, *19*(3), 273–288. https://doi.org/10.1080/15472450.2013.868284

Perez, A. J., Dominguez, L. D., Rubiales, A. J., & Lotito, P. A. (2015). Optimización de matrices origen-destino estimadas a partir de datos georeferenciados en redes sociales. *13º Simposio Argentino de Investigación Operativa*, 47–56.

Picornell, M., Ruiz, T., Lenormand, M., Ramasco, J. J., Dubernet, T., & Frías-Martínez, E. (2015). Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation*, *42*(4), 647–668. https://doi.org/10.1007/s11116-015-9594-1

Popescu, A., & Pennacchiotti, M. (2011). Dancing with the Stars, NBA Games, Politics:

An Exploration of Twitter Users' Response to Events. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 594–597. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2866/3233

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies*, *75*, 197–211. https://doi.org/10.1016/j.trc.2016.12.008

Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management*, *57*, 202–212. https://doi.org/10.1016/j.tourman.2016.06.006

Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: analysing cities using the space – time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, *36*(5), 824–836. https://doi.org/10.1068/b34133t

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, *6*(2), 1–27. https://doi.org/10.1145/1689239.1689243

Ren, F., & Kwan, M.-P. (2007). Geovisualization of Human Hybrid Activity-Travel Patterns. *Transactions in GIS*, *11*(5), 721–744. https://doi.org/10.1111/j.1467-9671.2007.01069.x

Resch, B., Zipf, A., Beinat, E., Breuss-Schneeweis, P., & Boher, M. (2012). *Towards the Live City-Paving the Way to Real-time Urbanism*. Retrieved from https://www.researchgate.net/publication/235708617

Rodrigue, J.-P., Comtois, C., & Slack, B. (2016). *The Geography of Transport Systems*. Routledge. https://doi.org/10.4324/9781315618159

Romanillos, G., & Zaltz Austwick, M. (2016). Madrid cycle track: visualizing the cyclable city. *Journal of Maps*, *12*(5), 1218–1226. https://doi.org/10.1080/17445647.2015.1088901

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *WWW2010*. Retrieved from

http://mecab.sourceforge.net/

Saladié, Ò., & Jurado, J. (2015). La movilidad en el Campus Vila-seca de la URV: propuestas para una movilidad más sostenible. *Investigaciones Geográficas*, (64), 163–182. https://doi.org/10.14198/INGEO2015.64.10

Salas-Olmedo, M. H., Moya-Gómez, B., García-Palomares, J. C., & Gutiérrez, J. (2018). Tourists' digital footprint in cities: Comparing Big Data sources. *Tourism Management*, *66*, 13–25. https://doi.org/10.1016/j.tourman.2017.11.001

Salas-Olmedo, M. H., & Rojas Quezada, C. (2017). The use of public spaces in a medium-sized city: from Twitter data to mobility patterns. *Journal of Maps*, *13*(1), 40–45. https://doi.org/10.1080/17445647.2017.1305302

Saura, J. R., & Bennett, D. R. (2019). A three-stage method for data text mining: Using UGC in business intelligence analysis. *Symmetry*, *11*(4). https://doi.org/10.3390/sym11040519

Schwab, K. (2017). *The fourth industrial revolution*. Crown Business.

Schwanen, T. (2017). Geographies of transport II: Reconciling the general and the particular. *Progress in Human Geography*, *41*(3), 355–364. https://doi.org/10.1177/0309132516628259

Schwanen, T., & Kwan, M. P. (2008). The Internet, mobile phone and space-time constraints. *Geoforum*, *39*(3), 1362–1377. https://doi.org/10.1016/j.geoforum.2007.11.005

Schweitzer, L. (2014). Planning and social media: A case study of public transit and stigma on twitter. *Journal of the American Planning Association*, *80*(3), 218–238. https://doi.org/10.1080/01944363.2014.980439

Segerberg, A., & Bennett, W. L. (2011). Social media and the organization of collective action: Using twitter to explore the ecologies of two climate change protests. *Communication Review*, *14*(3), 197–215. https://doi.org/10.1080/10714421.2011.597250

Seguí-Pons, J. M., Ruiz, M., & Luna, M. (2013). Movilidad y transportes en el acceso al Campus de la Universitat de les Illes Balears: una perspectiva de género. *XIII Congreso de Geógrafos Españoles.*, 685–695.

Shannon, T., Giles-Corti, B., Pikora, T., Bulsara, M., Shilton, T., & Bull, F. (2006). Active commuting in a university setting: Assessing commuting habits and potential for modal change. *Transport Policy*, *13*(3), 240–253. https://doi.org/10.1016/j.tranpol.2005.11.002

Shaw, S. L., Yu, H., & Bombom, L. S. (2008). A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS*, *12*(4), 425–441. https://doi.org/10.1111/j.1467-9671.2008.01114.x

Shekhar, S., Gunturi, V., Evans, M. R., & Yang, K. (2012). Spatial big-data challenges intersecting mobility and cloud computing. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access - MobiDE '12* (Vol. 1, p. 1). New York, New York, USA: ACM Press. https://doi.org/10.1145/2258056.2258058

Shelton, T. (2017). Spatialities of data: mapping social media 'beyond the geotag.' *GeoJournal*, *82*(4), 721–734. https://doi.org/10.1007/s10708-016-9713-3

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, 198–211. https://doi.org/10.1016/j.landurbplan.2015.02.020

Shen, Yao, & Karimi, K. (2016). Urban function connectivity: Characterisation of functional urban streets with social media check-in data. *Cities*, *55*, 9–21. https://doi.org/10.1016/j.cities.2016.03.013

Shen, Yue, Kwan, M.-P., & Chai, Y. (2013). Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. *Journal of Transport Geography*, *32*, 1–11. https://doi.org/10.1016/j.jtrangeo.2013.07.007

Simpson, R. (2017). Android overtakes Windows for first time. Retrieved April 20, 2018, from http://gs.statcounter.com/press/android-overtakes-windows-for-first-time

Soria-Lara, J. A., Marquet, O., & Miralles-Guasch, C. (2017). The influence of location, socioeconomics, and behaviour on travel-demand by car in metropolitan university campuses. *Transportation Research Part D: Transport and Environment*, *53*, 149–160. https://doi.org/10.1016/j.trd.2017.04.008

Soria-Lara, J. A., Miralles-Guasch, C., & Marquet, O. (2017). The influence of lifestyle and built environment factors on transport CO2 emissions: the case study of Autonomous University of Barcelona. *ACE: Architecture, City and Environment*. https://doi.org/10.5821/ace.12.34.4756

Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. *Transactions in GIS*, *19*(6), 809–834. https://doi.org/10.1111/tgis.12132

Steiger, E., Lauer, J., & Ellersiek, T. (2014). Towards a framework for automatic geographic feature extraction from Twitter. *Eighth International Conference on Geographic Information Science*. Retrieved from http://koenigstuhl.geog.uni-heidelberg.de/publications/2014/Steiger/GISCIENCE_Steiger_et_al_2014_geographicfeatureextraction.pdf

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, *30*(9), 1694–1716. https://doi.org/10.1080/13658816.2015.1099658

Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, *54*, 255–265. https://doi.org/10.1016/j.compenvurbsys.2015.09.007

Stephens, M., & Poorthuis, A. (2015). Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems*, *53*, 87–95. https://doi.org/10.1016/j.compenvurbsys.2014.07.002

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, *25*(11), 1737–1748. https://doi.org/10.1080/13658816.2011.604636

Tao, S., Rohde, D., & Corcoran, J. (2014). Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. *Journal of Transport Geography*, *41*(December), 21–36. https://doi.org/10.1016/j.jtrangeo.2014.08.006

Tong, L., Zhou, X., & Miller, H. J. (2015). Transportation network design for

maximizing space–time accessibility. *Transportation Research Part B: Methodological*, *81*, 555–576. https://doi.org/10.1016/j.trb.2015.08.002

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., & González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, *58*, 162–177. https://doi.org/10.1016/j.trc.2015.04.022

Twitter. (2019). Tweet objects. Retrieved October 22, 2019, from https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

Versichele, M., Neutens, T., Delafontaine, M., & Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, *32*(2), 208–220. https://doi.org/10.1016/j.apgeog.2011.05.011

Volosin, S. E., Paul, S., Pendyala, R. M., Livshits, V., & Maneva, P. (2013). Activity-Travel characteristics of a large university population. *Urban Transportation Data and Information Systems*, *7*, 1–19.

Wachowicz, M., & Liu, T. (2016). Finding spatial outliers in collective mobility patterns coupled with social ties. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2016.1144887

Wang, Q., Phillips, N. E., Small, M. L., & Sampson, R. J. (2018). Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(30), 7735–7740. https://doi.org/10.1073/pnas.1802537115

Wang, X., Khattak, A., Son, S., & Agnello, P. (2012). What can be learned from analyzing university student travel demand? *Transportation Research Record: Journal of the Transportation Research Board*, *2322*, 129–137. https://doi.org/10.3141/2322-14

Weng, J., Yao, Y., Leonardi, E., & Lee, F. (2011). Event Detection in Twitter. *Proceedings of the Fifth International Conference on Weblogs and Social Media*.

Whalen, K. E., Páez, A., & Carrasco, J. A. (2013). Mode choice of university students commuting to schooland the role of active travel. *Journal of Transport Geography*,

31, 132–142. https://doi.org/10.1016/j.jtrangeo.2013.06.008

Williams, S. A., Terras, M., & Warwick, C. (2013). What people study when they study Twitter. *Journal of Documentation*, *69*(3), 1–74. https://doi.org/10.1108/JD-03-2012-0027

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-urban human mobility and activity transition: evidence from social media check-in data. *PloS One*, *9*(5), e97010. https://doi.org/10.1371/journal.pone.0097010

Xia, F., Yang, L. T., Wang, L., & Vinel, A. (2012). Internet of Things. *International Journal of Communication Systems*, *25*(9), 1101–1102. https://doi.org/10.1002/dac.2417

Xu, Y., & González, M. C. (2017). Collective benefits in traffic during mega events via the use of information technologies. *Journal of The Royal Society Interface*, *14*(129), 1–10. https://doi.org/10.1098/rsif.2016.1041

Yin, J., Soliman, A., Yin, D., & Wang, S. (2017). Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science*, *31*(7), 1293–1313. https://doi.org/10.1080/13658816.2017.1282615

Yin, L., Shaw Shih-Lung, S. L., & Yu, H. (2011). Potential effects of ICT on face-to-face meeting opportunities: A GIS-based time-geographic approach. *Journal of Transport Geography*, *19*(3), 422–433. https://doi.org/10.1016/j.jtrangeo.2010.09.007

Yu, H., & Shaw, S. (2008). Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. *International Journal of Geographical Information Science*, *22*(4), 409–430. https://doi.org/10.1080/13658810701427569

Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. *Tourism Management Perspectives*, *10*, 27–36. https://doi.org/10.1016/j.tmp.2014.01.001

Zhañay, B. A., Cordero, G. O., Cordero, M. O., & Urigüen, M.-I. A. (2019). A Text Mining Approach to Discover Real-Time Transit Events from Twitter (pp. 155–169). Springer, Cham. https://doi.org/10.1007/978-3-030-02828-2_12

Zhang, P., Zhou, J., & Zhang, T. (2017). Quantifying and visualizing jobs-housing

balance with big data: A case study of Shanghai. *Cities*, *66*, 10–22. https://doi.org/10.1016/j.cities.2017.03.004

Zhang, S., & Feick, R. (2016). Understanding public opinions from geosocial media. *ISPRS International Journal of Geo-Information*, *5*(6). https://doi.org/10.3390/ijgi5060074

Zhao, F., Ghorpade, A., Pereira, F. C., Zegras, C., & Ben-Akiva, M. (2015). Quantifying mobility. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers - UbiComp '15* (pp. 1039–1044). New York, New York, USA: ACM Press. https://doi.org/10.1145/2800835.2800957

Zhen, F., Cao, Y., Qin, X., & Wang, B. (2017). Delineation of an urban agglomeration boundary based on Sina Weibo microblog 'check-in' data: A case study of the Yangtze River Delta. *Cities*, *60*, 180–191. https://doi.org/10.1016/j.cities.2016.08.014

Zhou, J. (2014). From better understandings to proactive actions: Housing location and commuting mode choices among university students. *Transport Policy*, *33*, 166–175. https://doi.org/10.1016/j.tranpol.2014.03.004

Zhou, X., & Xu, C. (2017). Tracing the Spatial-Temporal Evolution of Events Based on Social Media Data. *ISPRS International Journal of Geo-Information*, *6*(3), 88. https://doi.org/10.3390/ijgi6030088

tGIS
transporte infraestructuras y territorio
Grupo de Investigación de la Universidad Complutense de Madrid

Esta tesis doctoral está enmarcada en el grupo de investigación tGIS de la Universidad Complutense de Madrid. La tesis ha contado con la financiación del Ministerio de Educación, Cultura y Deporte a través de la beca FPU (Programa FPU AP2015-0147), las ayudas a la movilidad para estancias breves y traslados temporales de beneficiarios FPU para la realización de una estancia breve en la Universidad VSB Técnica de Ostrava (República Checa), del Ministerio de Economía, Industria y Competitividad (MINECO) y el Fondo de Desarrollo Regional Europeo (ERDF) (Proyecto TRA2015-65283-R), de la Comunidad de Madrid (SOCIALBIGDATA-CM, S2015/HUM-3427), y del Ministerio de Ciencia, Innovación y Universidades y el Fondo Regional Europeo de Desarrollo (Proyecto DynMobility, RTI2018-098402-B-I00).